HIERARCHICAL HUMAN ACTIVITY RECOGNITION WITH FUSION OF AUDIO AND MULTIPLE INERTIAL SENSOR MODALITIES

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

TUĞÇE ALARA YILMAZ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

FEBRUARY 2022

Approval of the thesis:

HIERARCHICAL HUMAN ACTIVITY RECOGNITION WITH FUSION OF AUDIO AND MULTIPLE INERTIAL SENSOR MODALITIES

submitted by **TUĞÇE ALARA YILMAZ** in partial fulfillment of the requirements for the degree of **Master of Science in Computer Engineering Department, Middle East Technical University** by,

Prof. Dr. Halil Kalıpçılar
Dean, Graduate School of **Natural and Applied Sciences** _____

Prof. Dr. Halit Oğuztüzün
Head of Department, **Computer Engineering** _____

Prof. Dr. Adnan Yazıcı
Supervisor, **Computer Engineering, METU** _____

Examining Committee Members:

Assoc. Prof. Dr. Sinan Kalkan
Computer Engineering, METU _____

Prof. Dr. Adnan Yazıcı
Computer Engineering, METU _____

Assoc. Prof. Dr. Enver Ever
Computer Engineering, METU Northern Cyprus Campus _____

Date: 08.02.2022

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Surname: Tuğçe Alara Yılmaz

Signature :

ABSTRACT

HIERARCHICAL HUMAN ACTIVITY RECOGNITION WITH FUSION OF AUDIO AND MULTIPLE INERTIAL SENSOR MODALITIES

Yılmaz, Tuğçe Alara
M.S., Department of Computer Engineering
Supervisor: Prof. Dr. Adnan Yazıcı

February 2022, 97 pages

People perform a wide variety of activities every day. Systems that can automatically distinguish these activities, i.e. human activity recognition models, have improved markedly, especially in the last decade. Deep learning is demonstrating increasingly promising outcomes in overcoming the problem of human activity detection as technology improves at a rapid pace. However, validating activity recognition in real-world situations is critical for practical solutions that work in natural contexts. Establishing systems that could achieve automatic activity recognition with real-life settings such as the devices that people use every day naturally and the environment they live in, might require lots of computational complexity. A lightweight neural network model is adopted for this purpose, one that could run swiftly even in the background process without taking up a lot of space when embedded into smartphones. Four inertial sensory data are represented in color coded image form and fused with three channelled audio data image representation to perform recognition task. The resulting fusion images allow rapid recognition performance because the size of the each image is so small. This thesis also underlines that audio sensor data, which require considerably bigger window sizes for identification on their own, improve

automated recognition performance when used in conjunction with inertial sensors, even when divided into small window sizes to interact with other sensors simultaneously. In addition, this thesis provides a strategy that helps the computer better discern real-world behaviors by introducing activities, contexts, and placements in a hierarchical manner to perform accurate activity recognition. By merging auditory images with inertial color coded images, representing multiple activity pairs with hierarchical groups according to activity-context-placement, and employing a lightweight model, high accurate recognition performance score competitive to state-of-art, nearly 91% success rate is achieved. We believe that this research could be classified as a "quality of experience" because it presents a lightweight model that could be used to predict behavior of individual by data collected from devices such as smartphone and smartwatch that everyone uses each day naturally.

Keywords: human activity recognition, deep learning, inertial sensor, audio sensor

ÖZ

SES VE ÇOKLU ATALET SENSÖRÜ MODALİTELERİ FÜZYONU İLE HİYERARŞİK İNSAN AKTİVİTESİ TANIMA

Yılmaz, Tuğçe Alara
Yüksek Lisans, Bilgisayar Mühendisliği Bölümü
Tez Yöneticisi: Prof. Dr. Adnan Yazıcı

Şubat 2022 , 97 sayfa

İnsanlar her gün çok çeşitli aktiviteler gerçekleştirirler. Bu faaliyetleri otomatik olarak ayırt edebilen sistemler, yani insan faaliyeti tanıma modelleri, özellikle son on yılda önemli ölçüde iyileşmiştir. Derin öğrenme, teknoloji hızla geliştikçe insan etkinliği tespiti sorununun üstesinden gelmede giderek daha fazla umut vaat eden sonuçlar ortaya koyuyor. Bununla birlikte, gerçek dünya durumlarında etkinlik tanımayı doğrulamak, doğal bağlamlarda çalışan pratik çözümler için kritik öneme sahiptir. İnsanların her gün doğal olarak kullandıkları cihazlar ve yaşadıkları ortam gibi gerçek hayat ayarlarıyla otomatik aktivite tanımayı gerçekleştirebilecek sistemler kurmak, çok fazla hesaplama karmaşıklığı gerektirebilir. Bu amaç için, akıllı telefonlara yerleştirildiğinde çok fazla yer kaplamadan arka planda bile hızla çalışabilen hafif bir sinir ağı modeli benimsenmiştir. Dört atalet sensör verisi, renk kodlu görüntü biçiminde temsil edilir ve tanıma görevini gerçekleştirmek için üç kanallı ses verisi görüntü temsili ile birleştirilir. Elde edilen füzyon görüntüleri, her görüntünün boyutu çok küçük olduğu için hızlı tanıma performansı sağlar. Bu tez ayrıca, kendi başlarına tanımlama için önemli ölçüde daha büyük pencere boyutları gerektiren ses sensörü

verilerinin, diğer sensörlerle aynı anda etkileşim kurmak için küçük pencere boyutlarına bölündüğünde bile, eylemsiz sensörlerle birlikte kullanıldığında otomatik tanıma performansını iyileştirdiğinin altını çizer. Ek olarak, bu tez, doğru etkinlik tanıma gerçekleştirmek için etkinlikleri, bağlamları ve yerleşimleri hiyerarşik bir şekilde sunarak bilgisayarın gerçek dünya davranışlarını daha iyi ayırt etmesine yardımcı olan bir strateji sağlar. İşitsel görüntüleri eylemsiz renk kodlu görüntülerle birleştirerek, etkinlik-bağlam-yerleştirmeye göre hiyerarşik gruplarla çoklu etkinlik çiftlerini temsil ederek ve hafif bir model kullanarak, son teknoloji ile rekabet eden yüksek doğrulukta tanıma performansı puanı, yaklaşık 91% başarı sonucu elde edilir. Herkesin her gün doğal olarak kullandığı akıllı telefon ve akıllı saat gibi cihazlardan toplanan verilerle bireyin davranışını tahmin etmek için kullanılabilecek hafif bir model sunduğu için bu araştırmanın bir "deneyim kalitesi" olarak sınıflandırılabileceğine inanıyoruz.

Anahtar Kelimeler: insan aktivitesi tanıma, derin öğrenme, atalet sensörü, ses sensörü

To my warm, kind, beloved family

ACKNOWLEDGMENTS

First and foremost, I'd want to thank my wonderful advisor Prof. Dr. Adnan Yazıcı for illuminating my way, throughout my short thesis period with him. I am grateful to him for his guidance and support that helped me during this thesis, as well as his warmth, friendliness and constantly positive attitude.

I'd want to express my gratitude to Assoc. Prof Dr. Enver Ever, with whom I had the pleasure of meeting and working throughout the process. I would also like to thank Hakan Yekta Yatbaz, who help me to take the thesis much further by making pinpoint detections on every problem I encountered during this period.

My grateful thanks to Assoc. Prof. Dr. Sinan Kalkan for his participation in the thesis period and mentoring me about Deep Learning.

I would like to express my gratitude to my one and only family, my beloved mother, father and my dearest little brother Engin Yılmaz, who constantly encouraged and supported me in this process, as in every moment of my life, and made me feel their warmth, kindness and sincere support during the difficult times.

I'd like to thank my other half, Hasan Ali Duran, who has been one of my biggest supporters during this intense phase and life by understanding me better than I do. I am grateful for his presence in my life and for his constant support in all situations as well as in this thesis period.

Special thanks to my dear best friend Seda Polat, who always back me up though my most of the life since I was six years old and uplifting my moral during this process.

My thanks to my classmates whom I had the opportunity to meet during my graduate studies, it was a joy to share ideas with them. It was a long journey for all of us, but we all made it. Finally, I would like to thank my colleagues for their mental support during this process.

TABLE OF CONTENTS

ABSTRACT v

ÖZ vii

ACKNOWLEDGMENTS x

TABLE OF CONTENTS xi

LIST OF TABLES xiv

LIST OF FIGURES xvi

LIST OF ABBREVIATIONS xviii

CHAPTERS

1 INTRODUCTION 1

1.1 Human Activity Recognition in Real-World 1

1.2 Objectives of the Thesis 2

1.3 Contributions of the Thesis 3

1.4 The Outline of the Thesis 5

2 BACKGROUND 7

2.1 Machine Learning 7

2.2 Deep Learning 11

2.2.1 Convolutional Neural Network 11

2.2.2 Recurrent Neural Network 13

2.2.3 Autoencoder 14
2.2.4 Restricted Boltzmann Machine 14
2.3 Performance Evaluation in ML & DL 15
2.3.1 Cross-Validation Methods 15
2.3.2 Performance Metrics 16
2.4 Related Work 17
2.4.1 Human Activity Recognition (HAR) 17
2.4.2 Datasets used for HAR 17
2.4.2.1 The ExtraSensory Dataset Details 17
2.4.3 HAR Studies 20
2.4.3.1 Sensor Based HAR Studies 20
2.4.3.2 Vision Based HAR Studies 26
2.5 Audio Processing 29
3 PROPOSED FRAMEWORK FOR HUMAN ACTIVITY RECOGNITION . 31
3.1 The Framework for Hierarchical Human Activity Recognition 31
3.2 Input Generation For Human Activity Recognition Task 32
3.2.1 Data Preprocessing 32
3.2.2 Image Generation 34
3.2.2.1 Inertial Sensor Images 36
3.2.2.2 Audio Images 37
3.3 Convolutional Neural Network Model 38
4 EXPERIMENTS AND EVALUATION 43
4.1 Hierarchical Evaluation based on Main Group Independency 44

4.1.1 Holdout Cross Validation Results 45
4.1.2 5-Fold Cross Validation Results 51
4.2 Hierarchical Evaluation based on Main Group Dependency 60
4.2.1 Holdout Cross Validation Results 61
4.2.2 5-Fold Cross Validation Results 71
4.3 Discussion 80
5 CONCLUSION 87
REFERENCES 89

LIST OF TABLES

TABLES

Table 2.1 Public Datasets 18

Table 2.2 Abbreviaton and Activity Matching 19

Table 2.3 Sensor-Based HAR Studies 21

Table 2.4 Sensor-Based HAR Studies (continued) 22

Table 2.5 Vision-Based HAR Studies 27

Table 4.1 Holdout Results of Primary Physical Activities 46

Table 4.2 Holdout Results of Placement Independent Contexts 47

Table 4.3 Holdout Results of Placement Dependent Contexts 49

Table 4.4 Holdout Results of Placement Dependent Contexts (continued) . . . 50

Table 4.5 5-Fold Results of Primary Physical Activities 52

Table 4.6 5-Fold Results of Primary Physical Activities (continued) 53

Table 4.7 5-Fold Results of Placement Independent Contexts 55

Table 4.8 5-Fold Results of Placement Independent Contexts (continued) . . . 56

Table 4.9 5-Fold Results of Placement Dependent Contexts 57

Table 4.10 5-Fold Results of Placement Dependent Contexts (continued) 58

Table 4.11 Holdout Results of Primary Physical Activity Results 62

Table 4.12 Average Holdout Results of Placement Independent Context Group Testing 63

Table 4.13 Average Holdout Results of Placement Dependent Context Group Testing 67

Table 4.14 5-Fold Results of Primary Physical Activities 72

Table 4.15 5-Fold Results of Primary Physical Activities (continued) 73

Table 4.16 Average 5-Fold Results of Placement Independent Context Group Testing 74

Table 4.17 Average 5-Fold Results of Placement Dependent Context Group Testing 77

Table 4.18 Studies using the real-world ExtraSensory dataset and results 84

Table 4.19 Average Memory Space of Model 85

LIST OF FIGURES

FIGURES

Figure 2.1 Basic neuron structure 10

Figure 2.2 Basic CNN structure 12

Figure 2.3 Pooling Operation and Types 13

Figure 2.4 Audio Signal Processing steps 30

Figure 3.1 Proposed framework for hierarchical human activity recognition 32

Figure 3.2 Data Clustering Flow 34

Figure 3.3 Inertial sensory fusion image of Lying Down Activity with Units 36

Figure 3.4 Inertial images of sample classes with hierarchy 36

Figure 3.5 Audio Feature Images 37

Figure 3.6 Accuracy vs Epoch Graphs of each Learning Rate 40

Figure 3.7 CNN structure used in this work 42

Figure 4.1 Hierarchically User-Defined Activities and Test Groups - Main Group Independent 44

Figure 4.2 PPA Main Group Independent F1 scores per activity - Holdout . 45

Figure 4.3 PIC Main Group Independent F1 scores per context - Holdout . . 48

Figure 4.5 PPA Main Group Independent F1 scores per activity - 5-Fold . . 54

Figure 4.6 PIC Main Group Independent F1 scores per context - 5-Fold . . 56

Figure 4.8 Hierarchically User-Defined Activities and Test Groups - Main Group Dependent 60

Figure 4.9 PPA Main Group Dependent F1 scores per activity - Holdout . . 61

Figure 4.10 PIC Main Group Dependent F1 scores per context - Holdout . . 64

Figure 4.11 H2 Test Group - Placement Independent Context-3 Confusion Matrices, iter 4 (a) Inertial & MSPC fusion (b) Inertial & MFCC fusion (c) Inertial (d) MSPC (e) MFCC - Holdout 66

Figure 4.12 H3 Test Group - Placement Independent Context-3 Confusion Matrices, iter 2 (a) Inertial & MSPC fusion (b) Inertial & MFCC fusion (c) Inertial (d) MSPC (e) MFCC - Holdout 69

Figure 4.14 PPA Main Group Dependent F1 scores per activity - 5-Fold . . . 71

Figure 4.15 PIC Main Group Dependent F1 scores per context - 5-Fold . . . 75

Figure 4.16 H2 Test Group - Placement Dependent Context-3 Confusion Matrices, Fold 1 (a-b) MFCC, Inertial & MFCC (c-d) MSPC, Inertial&MSPC, (e-f) LMSPC, Inertial&LMSPC, - 5-Fold 76

Figure 4.18 H3 Test Group - Placement Dependent Context-3 Confusion Matrices, Fold 3 (a-b) MFCC, Inertial & MFCC (c-d) MSPC, Inertial&MSPC, (e-f) LMSPC, Inertial&LMSPC, - 5-Fold 79

LIST OF ABBREVIATIONS

2D	2 Dimensional
3D	3 Dimensional
A	Accuracy
ADAM	Adaptive Moment Estimation
AI	Artificial Intelligence
ANN	Artificial Neural Network
BA	Balanced Accuracy
CNN	Convolutional Neural Network
DBN	Deep Belief Network
DCT	Discrete Cosine Transform
DFT	Discrete Fourier Transform
DL	Deep Learning
DT	Decision Tree
FC	Fully Connected
GRU	Gated Recurrent Unit
HAR	Human Activity Recognition
IoT	Internet of Things
k-NN	K-Nearest Neighbour
LMSPC	Log Mel Spectrogram
LOO	Leave One Day Out
LOO	Leave One Out
LOO	Leave One Subject Out
LSTM	Long Short-Term Memor
Mac-F1	Macro F1

Mic-F1	Micro F1
ML	Machine Learning
MFCC	Mel-Frequency Cepstrum
MSPC	Mel Spectrogram
PDC	Placement Dependent Context
PIC	Placement Independent Context
PPA	Primary Physical Activity
PS	Power Spectrum
RBM	Restrictive Boltzmann Machine
ReLU	Rectified Linear Unit
RF	Random Forest
RGB	Red-Green-Blue
RGB-D	Red-Green-Blue-Depth
RMSProp	Root Mean Square Propagation
RNN	Recurrent Neural Networks
SGD	Stochastic Gradient Descent
SVM	Support Vector Machine
W-F1	Weighted F1

CHAPTER 1

INTRODUCTION

1.1 Human Activity Recognition in Real-World

Human activity recognition tasks take advantage of assistive technology to provide knowledge of what activity is performed by users by constructing a pattern. Recognizing human activities via connected sensors and widely available wearable computing, sometimes known as the Internet of Things (IoT), is an interesting topic with numerous challenges. Some of those challenges are complexity of natural activities, inter and intra changeability of natural behavior and keeping the balance preserving privacy issue and research.

Human activity recognition applications are used to tackle problems in variety of fields such as elderly care [49], sports [50], smart home gain popularity. Automatic detection with a robust algorithm makes the process more frequent, objective and effortless. In order for these applications to be successful on a broad scale, the context recognition solution must be inconspicuous and operate without needing the user to change their behavior. It's critical that research reflects real-world scenarios in which such solutions would be used by people in their natural environment with the help of every-day devices. The fact that this operation takes place in the background while individuals using their own devices without any guidance prevents it from being a burdensome process for people. Simultaneously, to be effective in the field of human activity recognition, an approach that is not specific to a person but can forecast the daily behavior of the general population is required. By offering a generic framework, this thesis seeks to distinguish human activities or behaviors that various people conduct on a regular basis with varied habits. It achieves this purpose by combining

numerous sensor modalities.

1.2 Objectives of the Thesis

The primary goal of this research is to offer a model that examines human behavior in a hierarchical fashion using multiple images coded from sensors implanted in devices that could be utilized by everyone in their daily lives in conjunction with technology. To achieve so, color coded inertial sensor images and audio images are fused with small image representation to be used during activity prediction. Simultaneously, the effects of audio and inertial sensors on activity recognition are studied independently, while the impact of combining multiple modalities on basic performance is explored. The multi-label problem is efficiently solved using a hierarchical way to define physical activity, context, and phone placement.

The following tasks are examined in order to fulfill the provided objectives:

- Examining the existing studies of human behavioral context recognition, for both sensor based and vision based
- Studying over common deep learning and machine learning approaches
- Examining audio sensory features and converting audio sensor data
- Converting multiple sensory data to color coded images and audio data to images with three channel
- Proposing hierarchic method to combine physical activity, context and device placement
- Working with a deep neural network model used multiple embedded sensor modalities from smartphone and smartwatch
- Conducting evaluation of proposed system

1.3 Contributions of the Thesis

The research emphasizes on using inertial and audio sensor modalities' fusion in a form that works well with a lightweight model. Model could predict people's natural behavioral contexts in their daily living environment and with the items they use in daily-base without being recorded in a controlled environment, in hierarchical way by predicting both physical activity and behavioral context. In order to achieve this, it shows the effect of combined different sensor modalities on human activity recognition by fusing the sensor with the features obtained from the audio sensor, as well as the sensors such as accelerometer and gyroscope, which are widely used in human activity recognition.

Many studies in the existing literature require individuals to complete prescribed tasks, which resulted in non-natural behavior. In order to reveal the effect of utilizing different sensor modalities while predicting behavior, a large real-life dataset is used. This data collection, which is compiled from the daily activities of a large number of participants, has far more activities than those reported in the literature. Simultaneously, this data are only gathered from widely utilized technology gadgets, such as smartphones and smartwatches. There are no additional gadgets required for the model represented in this work to operate with, only smartphone and smartwatch, data of basic devices used by each person every day is enough to achieve promising results for different activity groupings. Because the records comprise actions that people conduct in their daily lives and because each person's behavior habits differ, the data should be submitted to particular processing in order to provide a general answer. Therefore, a data processing method based on simulating changing pattern of the sensor input distribution with standard deviation, for each different activity is presented in this work, for both inertial and audio sensor information. The data representation is in small image format, color coded by embedding X,Y,Z dimensions into R,G,B channels for inertial sensors and assigning each three channels as extracted feature data for audio sensors. During creation of images, sliding window approach is employed. One of the most crucial factors in activity recognition is the supplied window range. Small window sizes could be used by inertial sensors to distinguish activities, whereas wide frames are preferable for activity recognition from sound.

However, the window size in this study is maintained modest in order to see how well inertial and acoustic sensor fusions work together to differentiate activity.

People might perform multiple activities at once in real life, therefore, human activities are pre-defined within this work based on hierarchy, which indicated combination of multiple activities, based on depth level. First step of the hierarchy indicates a main physical activity such as "Walking" while second step declare context of human, for example "Shopping". Third step indicates smartphone placement, which gives cue about activity. The aim of this approach is to get the best results possible by comparing the effects of various physical activities or different phone placements reported simultaneously with a context. In this research, two alternative test procedures are used. All defined contexts (with or without phone placement) are tested with respect to the corresponding hierarchy depth in the first technique. The goal of this method is to create predictions based on a variety of activity combinations. Contexts with the same primary activity are examined with other contexts in their hierarchy layers as well as their own primary physical activity groups in the second procedure. The goal here is to discern distinct settings notably using auditory data, even though they have the same physical activity, and hence the same physical motion flow.

The study here also aims to achieve most accurate results by using minimized number of features gathered from sensors. Only three (X, Y, Z) dimensions are used for inertial data when generating color coded images, which are formed single pixel. For audio data, to maintain small image representation approach, MFCC and MSPC features are used only, which have 13 and 32 channels per each timestamp, respectively. Each channel represent a pixel in the audio image. Hyper parameters are tuned to achieve best possible results. Balanced accuracy and F1-scores are interpreted as efficiency units of the system uses imbalanced dataset. Overall performances are reported with BA while in context based performances are reported with F1 score. Compared to existing studies, model outperforms with balanced accuracy rate 91% for primary physical activity recognition. At second level, overall balanced accuracy score is measured as 82% while it is 85% at the third level of the hierarchical approach tree.

To summarize, the following key contributions are addressed in this thesis:

- A framework is represented which offers a generic inter-human solution to HAR. It processes data collected by devices each person use daily without causing additional burden. Furthermore, since the recordings are not scripted with respect to activity, the framework could be adapted into real-life solutions.
- A novel audio image representation which is incorporated to the existing solutions that use inertial sensors is introduced. Features are represented in 3 channelled image form to fuse audio sensory data with color coded images of inertial sensors.
- Activity, context and phone placements are modeled in a hierarchical manner to get better insight to the genuine human activity habits. This pre-definition of labels enables integrative context descriptions that contains a mix of activities, contexts, the environment, body position and other factors, while enabling model to operate multi-class classification method which actually indicates multi-label.
- A lightweight model is utilized that requires comparably little memory space and serves rapid prediction performance. Proposed model with data representation could be utilized in real-time activity monitoring solutions.
- Two methods are used to assess activity recognition performance. In the first method, described as main group independent, activity classes share same hierarchical level are examined as a whole group. In the second group, in addition to performing recognition within the same hierarchy, testing groups are constructed based on primary physical activity.

1.4 The Outline of the Thesis

There are five chapters in this thesis. In Chapter 2, the background material required to comprehend the rest of the work completed within the scope of this thesis is presented. The algorithms of machine learning and deep learning that are most commonly employed are summarized. The purpose of the literature review section is to

investigate the primary associated proposed techniques that are relevant to the applicative context.

In Chapter 3, the proposed model and design details are all thoroughly explained. In addition, one of the most significant contributions in this thesis, the explanation of how to analyze diverse sensor data and integrate them into the system in a different input form is described comprehensively.

In Chapter 4 the deep learning architecture provided for recognition with inertial and acoustic sensory inputs are thoroughly evaluated. Evaluation methods and observations are explained in detail. Furthermore, comparison is conducted with existing studies in the literature and this work.

The study is brought to a conclusion in Chapter 5 with summarizing observations.

CHAPTER 2

BACKGROUND

This chapter covers necessary background information about machine learning methods and deep learning architectures and used evaluation methods of those. Later, human activity recognition is explained briefly, common datasets are listed and related studies in the literature are examined in detail. Finally, the methods employed in the conversion of each feature of the audio signal, as well as the flow of this process, are detailed for the audio sensor data used in this study.

2.1 Machine Learning

Machine learning is an branch of Artificial Intelligence (AI) technology that allows systems to automatically learn and improve from experience without being explicitly designed. Machine Learning is concerned with the creation of computer programs that could access data and utilize it to learn on their own for given task.

There are several variances about defining types of Machine Learning but groups could be categorized with respect to their purpose. Based on their functionality, there are four main categories described as follows:

1. Supervised Learning: Supervised learning is a task of designing algorithms that can generate broad models and hypotheses using externally given situations to predict the destiny of future samples. The goal of supervised machine learning classification algorithms is to categorize data based on historical information. For simpler explanation, input and output pairs are mapped together [1]. There are two types of supervised learning, classification and regression.

(a) Classification: Classification is a process of categorizing a set of data into predetermined classes. Method aims to determine which category a given data belongs to by learning the relationship between input and output.

(b) Regression: In statistics, regression is the process of estimating an output value based on a collection of input values [4]. Regression models are used to predict dependent variable with respect to continuous set of independent variables.

2. Unsupervised Learning: Unsupervised learning is learning a task without having any information about correct outputs but by learning relationships among the perceptions and predicting future perceptions with the help of previous ones. There is no pre-assigned labeled data for unsupervised learning algorithm to work with [2].

3. Semi-supervised Learning: Chapelle et al. [3] defines semi-supervised learning as halfway between supervised learning and unsupervised learning, in a way that with unlabeled data, system is fed with some small portion of data that includes supervision information. Semi-supervised learning algorithms try to predict unlabeled data by understanding information from labeled data, under weak supervision.

4. Reinforcement Learning: Learning how to relate situations to behaviors in order to maximize a numerical reward or minimize the risk is known as reinforcement learning. Reinforcement learning, as opposed to other machine learning techniques, focuses on goal-directed learning via interaction [5].

In this thesis, classification is used to resolve the problem definition. Therefore, this section also covers widely-used classification algorithms.

Support vector machine (SVM) algorithm is a commonly used machine learning technique both used to achieve classification and regression tasks. It aims to find best hyperplane in N-dimensional space to separate data points from two classes. There could be more than one hyperplane that separates points, but SVM aims to find the one with maximum margin. Besides it is relatively memory efficient, algorithm is also effective when there the margin between point distributions is clearer.

K-Nearest Neighbors(k-NN) is another machine learning algorithm that is widely used since it is simple and easy-to-implement. Algorithm depends on the basic approach that similar things are commonly distributed over same or close space. For each point, algorithm tries to find belonging space by calculating its distances to samples already placed, based on a predefined distance function. To do so, it needs a distance vector and the number k.

Decision Tree (DT) which is a graph-based structure works like a flowchart that separating data points recursively with respect to feature-based conditions until reaching to leaf node, a class. Decision trees are easy to understand and they visualized graphically for non-experts to understand easily.

Random Forest (RF) represented by Leo Breiman is a set of individual decision trees where each tree is dependent to a vector generated by random sampling and each one has the same distribution across all other trees in the forest. Combining multiple models to make weak learner models has a significant vote. Aim of the RF algorithm is to receive more accurate prediction by advising most. It could be used for both classification and regression tasks.

Artificial Neural Network (ANN), forms the basis of frequently used deep learning algorithm is another machine learning algorithm used. ANN models are motivated by biological sciences, which investigate how real creatures' neuroanatomy has evolved to solve issues [7]. A conventional Artificial Neural Network is made up of several basic, linked processors known as neurons, each of which generates a series of real-valued activations [9]. In basic terms, a neuron accepts n inputs and plugs them into the equation $z = WX + \theta$, where X represents inputs $\{x_0, x_1 \dots x_n\}$, W represents weights of the inputs $\{w_0, w_j 1 \dots w_n\}$ and θ represents bias $\{\theta_0, \theta_1 \dots \theta_n\}$, where z is the output of the addition unit. Activation function $f_{act}(z)$ simulates the state of neuron and get output, $y = f_{act}(z)$ [8]. Figure 2.1 shows basic structure of neural network.

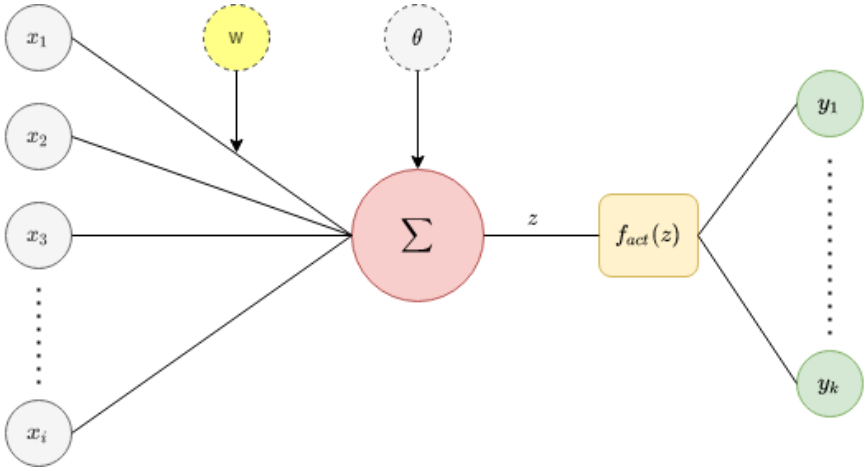


Figure 2.1: Basic neuron structure

The task of the activation function is to determine whether neurons will be activated based on the results from the weighted sum and bias. Activation functions are chosen specific to task to achieve. Most popular activation functions and their formula are given below:

1. Linear Activation:

$$\mathbf{f(x) = x} \tag{2.1}$$

2. Sigmoid Function:

$$\mathbf{f(x) = \frac{1}{1 + e^{-x}}} \tag{2.2}$$

3. Softmax Function:

$$\mathbf{f(x_i) = \frac{e^{x_i}}{\sum_{j=1}^{n} e^{x_j}}} \tag{2.3}$$

4. Hyperbolic Tangent (Tanh) Function:

$$\mathbf{f(x) = \frac{e^{2x} - 1}{e^{2x} + 1}} \tag{2.4}$$

5. Rectified Linear Unit(ReLU) Function:

$$\mathbf{f(x) = max(0, x)} \tag{2.5}$$

6. Leaky ReLU Function:

$$\mathbf{f(x) = max(0.1x, x)} \tag{2.6}$$

2.2 Deep Learning

Deep learning is a branch of machine learning that uses multilevel designs to acquire high-level abstractions from data [10]. It's a modern concept that's increasingly being used recently in conventional fields of Artificial Intelligence (AI) such as computer vision [11], natural language processing [12], transfer learning [13] and so on.

ANN algorithm is the basis of deep learning. A feedforward neural network means there are no cycling connections among the nodes. With this architecture, information flow is directed from input to output which is called forward propagation. In order to compute the gradient of the loss function regarding to weights, backpropagation allows the information from the scalar cost yielded by forward propagation in training phase to flow backward through the network[14]. This method is used to minimize the cost and maximize the efficiency. At this point, optimizer functions are also important because they aim the same. Optimizers are dependent to weights and biases of the model. Some well-known optimizer functions are Stochastic Gradient Descent(SGD), Mini-Batch Gradient Descent, Adaptive Moment Estimation (ADAM) and Root Mean Square Propagation (RMSProp).

There are many architectures such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Autoencoders, Restrictive Boltzmann Machines (RBM) etc. This section covers some neural network architectures used in deep learning literature to help better understanding of the work presented.

2.2.1 Convolutional Neural Network

CNNs are multilayered neural networks that use convolution operation to emphasis the importance of the information coming from closer neurons compared to further neurons. On contrary to neuron structure in ANN, neurons in CNN are structured in three dimensions, width-height which are spatial dimensions of input and depth which indicates the volume of the activation [16].

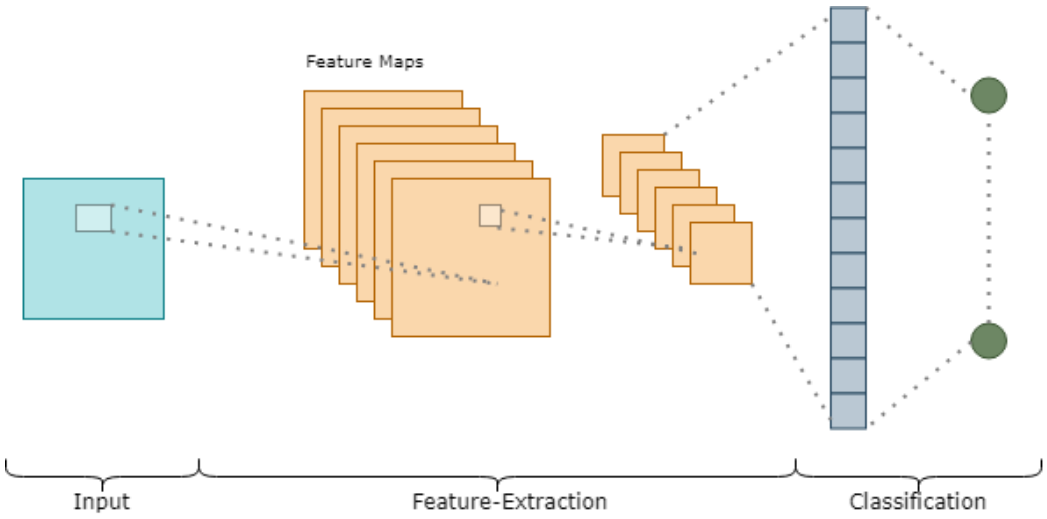


Figure 2.2: Basic CNN structure

A basic CNN composed of three parts; (1) input layer(s), (2) feature-extraction layers and (3) classification layer(s). Figure 2.2 shows basic structure. In the architecture, convolution layers are responsible for feature-extraction coming from previous neurons in hierarchy, which is handled by filters or in other name, kernels. Kernels in each convolution layer scan the input that has a typical grid-like structure [15] such as matrix and attempt to extract features that are relevant or explainable. Result of that operation is called feature map. As predicted, feature maps are dependent to kernel and input image, i.e., each time kernel is changed, feature maps are changed as well. Therefore, values of kernels are important. CNNs also help at this issue by finding optimized kernels at during learning phase. In order to operate convolution for each pixel with NxN convolution kernel first, each pixel in the image is scanned by using kernel. Then element-wise multiplication is applied with sub-pixels of the image and kernel values. Finally, results of each multiplication are written into the corresponding pixel.

Generally, a pooling layer follows convolution layer in CNN, which plays important role. It operates subsampling by reducing the size of the feature map, while eliminating less-informative points and preserving the critical information. They shrink dimension by a static kernel. They lower down computational cost since number of training parameters are decreased. There are different types of pooling; average, min-

imum and maximum. Figure 2.3 illustrates types and results.

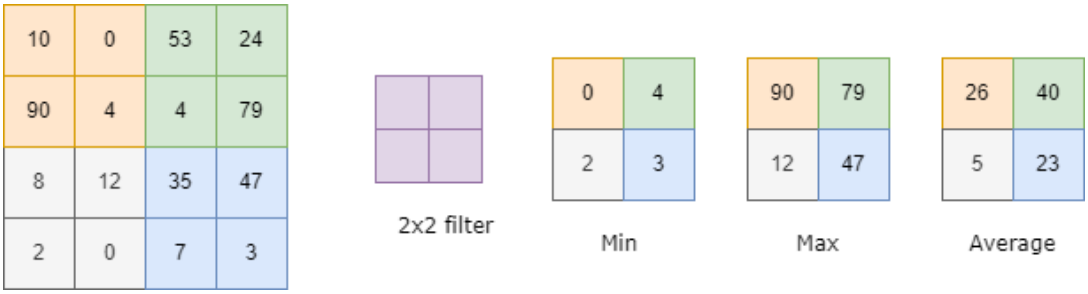


Figure 2.3: Pooling Operation and Types

After convolution and pooling layers have completed their job, final feature map is fed to classification layers, which are fully connected (FC) -or by other name, dense- layers to calculate class or regression scores. However, in cases where feature maps have more than one dimension, before they are fed directly, they are reshaped or flattened to one dimension.

Dropout is another layer type used in neural network models, it neglects some nodes during training by random choice. Dropout layers are used to regularize deep neural networks to prevent from overfitting which is basically a system learns features of given training data only, not the features of problem in general.

2.2.2 Recurrent Neural Network

Recurrent Neural Network (RNN) is an extension of traditional neural network architecture that has hidden states and allows previous experiences to be used as upcoming feature in those states, therefore, information is kept within the network. It has cyclic structure, which is the main difference between RNNs and regular feed-forward networks, it could be though as interconnected series networks. Well known application developed based on RNNs are Long Short-Term Memory (LSTM)[17] and Gated Recurrent Units (GRU) [18]. LSTMs are more advanced forms of RNNs that could solve the vanishing gradient problem, problem of architecture that could not update value of weight during training by permitting data to be preserved. GRUs are simpler

versions of LSTMs, which composed of gated units that control the flow of information within the them without any independent memory cells, and structurally, there are no output gate.

2.2.3 Autoencoder

Autoencoder is a type of neural networks that tries to learn a compact representation of input data by applying beneficial manipulations. Aim of an autoencoder is to recreate input data by neglecting insignificant data, by learning important features of while minimizing the error between input and output data. This NN architecture is composed of encoder and decoder parts. Encoding part interprets the input data and compress it. The output from encoding part is compressed data or bottleneck. Then decoder part attempts to reconstruct the input from compressed data. In cases when using only single autoencoder does not enough for proper feature extraction, Stacked Autoencoders where each hidden layer's output is coupled to the input of the next hidden layer are used.

2.2.4 Restricted Boltzmann Machine

Restricted Boltzmann Machines (RBM) are generative and non-deterministic neural network architectures that learn probability distribution that maximizes the likelihood. Structurally, they have restricted connections between their two layers of neurons, called visible and hidden. There are no connection within the same layer. Visible neurons translate each input with weight and bias and forwards to hidden neurons, this process is called forward-pass. Hidden neurons create activation with weights and overall bias and pass values for re-construction, which is called backward-pass. Model learns features after a few forward and backward passes. Deep Belief Networks (DBN) are generative models that combines multiple stacked RMBSs that each stacked model performs non-linear transformations and pass that output as input layer to consecutive RBM.

2.3 Performance Evaluation in ML & DL

2.3.1 Cross-Validation Methods

Learning algorithms try to reach a generalization by conducting various operations on a given data, then they try to understand whether the model performs consistent or not based on prediction accuracy results of unseen but similar data. When model learns noise or random fluctuations come along with information coming from seen data as concept, the problem called overfitting might occurred. A reliable model should be able to give high results when tested not only with the data supplied to it during the train, but also with data that it has never seen belongs to same task. To make sure model is reliable, cross-validation is applied. Cross-validation is basically resampling over data by splitting it with different techniques as follows:

- Holdout: It is the simplest cross-validation technique. All data is divided into two, train and test set, or three by adding a validation set, depending on the case. Division is done by ratio such as 70% train, 30% test.

- Leave-One-Out (LOO): Each data point is removed as test data while others are train data in this method. This process continues for each observation in the dataset. Average value gathered from results of testing per each iteration indicate overall model performance.

- Leave-One-Subject-Out (LOSO): Train and test sets are constructed with data of different users. Each subject's data is utilized as a test iteratively, while remaining users' data being used for training.

- Leave-One-Day-Out (LODO): Used in large time series datasets where data is divided into days with respect to timestamps.

- N-fold: In this technique, dataset is divided into N equal parts. While one part is split as test, remaining N-1 parts form training set. Each part N is put to the test iteratively.

2.3.2 Performance Metrics

To measure system performance, there are various methods. Depending on the factors such as the goal, sample distributions in the dataset, several metrics could be employed to measure model performance. In general view, Accuracy and F1 scores are the two most used metrics. Accuracy indicates the fraction of the correct predictions of the model of the all samples. On the other hand, F1 score is harmonic mean of sample ratio of correctly identified as positives to all actual-positives (recall) and sample ratio of correctly identified as positives to all predicted-positives (precision). Formulas of performance metrics are as follows:

1. Accuracy Score:
$$\mathbf{A} = \frac{\mathbf{TP} + \mathbf{TN}}{\mathbf{TP} + \mathbf{TN} + \mathbf{FP} + \mathbf{FN}} \tag{2.7}$$
2. Precision:
$$\mathbf{PRE} = \frac{\mathbf{TP}}{\mathbf{TP} + \mathbf{FP}} \tag{2.8}$$
3. Recall (Sensitivity):
$$\mathbf{REC} = \frac{\mathbf{TP}}{\mathbf{TP} + \mathbf{FN}} \tag{2.9}$$
4. Specificity:
$$\mathbf{SPE} = \frac{\mathbf{TN}}{\mathbf{TN} + \mathbf{FP}} \tag{2.10}$$
5. Balanced Accuracy Score:
$$\mathbf{BA} = \frac{\mathbf{REC} + \mathbf{SPE}}{\mathbf{2}} \tag{2.11}$$
6. F1 Score:
$$\mathbf{F1} = \frac{\mathbf{2} \cdot \mathbf{PRE} \cdot \mathbf{REC}}{\mathbf{PRE} + \mathbf{REC}} \tag{2.12}$$

where:

- TP: True Positives, prediction points to class which observation belongs.
- FP: False Positives, prediction points to class which observation doesn't belong.
- TN: True Negatives, prediction doesn't point class which observation does not belong.
- FN: False Negatives, prediction doesn't point class which observation belongs.

2.4 Related Work

2.4.1 Human Activity Recognition (HAR)

A human activity is defined as a particular stream of actions in a repetitive continuation. Human Activity Recognition solutions (HAR) aim to identify an activity based on its characteristics. However, an activity characteristic might change from one person to another because each person performs activities differently due to habits, personal preferences or health reasons. Besides such limitations, present systems offer promising outcomes recognizing human activities [20] [32].

In this work, sensor-based HAR task is performed. Generally, HAR applications use specific to certain type of sensor modalities such as body-worn sensors, objects sensors and ambient sensors [21].

2.4.2 Datasets used for HAR

Table 2.1 covers most used public datasets in HAR studies includes inertial sensor data and their details. Listed datasets are mostly based on sensor data are embedded sensors in smartphones and smartwatches.

2.4.2.1 The ExtraSensory Dataset Details

The ExtraSensory [25] dataset contains data from 60 individuals who performed natural behaviors in their own lives and in their own environment, and they documented these actions using the devices they use, without any external control (someone directing them what to do and when to do it). The term *in-the-wild* represents this approach. Data is collected from smartphones and smartwatches, while various sensor modalities are included, accelerometer (both phone and watch), gyroscope, magnetometer, audio, location and phone state (e.g. battery). All raw data sensor measurements are transformed into 225 features with feature extraction technique, which are combination of time domain features and frequency domain features. In this work, for x,y,z dimensions of inertial sensors and 13-channels of MFCC, *mean* time feature is used.

Table 2.1: Public Datasets

Reference	Dataset	Frequency	#Subject	#Activity	Sensor(s)
[22]	Opportunity	30 Hz	4	16	A,G,M,O,AM
[23]	UCI-HAR	50 Hz	30	6	A, G
[24]	MHealth	50 Hz	10	12	A,G,M,ECG
[25]	ExtraSensory	40 Hz	60	51	A,G,M,Aud
[26]	USC-HAD	100 Hz	14	12	A,G
[27]	WISDM	20 Hz	29	6	A
[28]	PAMAP2	100 Hz	9	23	A,G,M
[29]	REALDISP	50 Hz	17	33	A,G,M,4DQ
[30]	DSADS	25 Hz	8	19	A,G,M
[31]	HHAR	100Hz	9	6	A,G

*A:Accelerometer, G:Gyroscope, M:Magnetometer, Aud:Audio, AM:Ambient Sensor, O:Object Sensor, ECG:Electrocardiograph, 4DQ:4D Quaternions

There are no fixed timeframes for actions, as subjects are expected to record their natural behavior in their regular basis. Furthermore, each person might have different habits during an activity, for instance, someone could walk fast while other walks slowly. Because of this, the time of each activity is captured, as well as the quantity of context-based contributions that varies from person to person, in the dataset. It changes from person-based as well. Therefore, overall dataset become an imbalanced dataset. The efforts to prevent the imbalanced dataset from reducing performance for activities with fewer samples are explained in chapter 5.

ExtraSensory consists of 51 labels which many of them occur at the same time, makes it multi-label. Multi-label learning is concerned with data instances that are simultaneously associated with numerous class labels. Despite performance of existed studies with multi-label solutions increased, still for this study, this approach is not applicable. However, labels in the dataset could be subdivided into categories as activity (e.g. Walking), context/ (e.g. Watching TV), phone state (e.g. Phone on Table). This categorization is represented in hierarchical way in this study, where h1 labels imply physical activities (e.g. Lying Down), h2 labels indicate a context (e.g. Watching TV)

occur together with an physical activity and h3 labels are combination of a physical activity, a context and the state of phone (e.g. Phone on Table). During data creation, three-letter abbreviations are used to identify the labels of this hierarchical structure from each of the original labels, including physical activities <H1>, behavioral contexts <H2> and phone placements <H3>. Representation <H1>_<H2>_<H3> is used while defining new contexts for this study, refer Table 2.2.

Table 2.2: Abbreviaton and Activity Matching

Abbreviation	Activity Name	Type	Level at Hierarchy
LYD	Lying Down	Activity	1
STN	Standing	Activity	1
SIT	Sitting	Activity	1
WLK	Walking	Activity	1
RUN	Running	Activity	1
BIC	Bicycling	Activity	1
SLE	Sleeping	Context	2
SUR	Surfing the Internet	Context	2
WAT	Watching TV	Context	2
IAC	In a Car	Context	2
IAM	In a Meeting	Context	2
IND	Indoors	Context	2
OUT	Outside	Context	2
SHP	Shopping	Context	2
TLK	Talking	Context	2
EXE	Exercise	Context	2
PIH	Phone in Hand	Placement	3
POT	Phone on Table	Placement	3
PIP	Phone in Pocket	Placement	3
PIB	Phone in Bag	Placement	3

Because humans might run numerous activities at the same time, a hierarchical categorization of activities is essential. Furthermore, it is critical to combine activities in order to solve the multi-label problem without eliminating various labelled information, representing a new format that express same combination with single label approach. The activities listed above are chosen based on their sample contribution

range across the dataset and their mutual exclusivity with other user-defined activities at the same hierarchy level.

2.4.3 HAR Studies

Human activity recognition aims to model people's daily activity patterns in either controlled or uncontrolled settings. The main challenge HAR approaches face is the variety of activities. People perform numerous activities in daily life. Furthermore, each person intend to behave in different patterns while operating same activity due to habits. The number of models that accurately predict human behavior are rapidly grow, parallel to portable device usage. Still, behavior detection task faces struggles due to computational efficiency of devices. In general, regular and reliable data is needed to generalize and distinguish a behavior, but devices could perform excessive data with complex algorithms effectively. HAR studies are divided into two categories regarding feature data usage, such as sensor based and vision based.

2.4.3.1 Sensor Based HAR Studies

Sensor-based HAR has been used in a variety of practical applications, including smart home, healthcare and elderly monitoring. Furthermore, the rapid growth of wireless sensor networks has resulted in a vast volume of data being collected from a variety of sensors, including wearable sensors, object sensors, and environmental sensors [47]. Wearable sensors generally aim to understand human movement and behavior. Smartphones, smartwatches, bracelets are integrated with variable wearable sensors as accelerometer, gyroscope, magnetometer. Object sensors are used to detect movements of specific objects and correlate them with human behavior. Radio-frequency identifications are used to track and identify people and items in the Internet of Things (IOT) environment. Environmental sensors embedded in the environment are used to track changes in environmental factors as a result of physical activity. Radars, pressure sensors and temperature sensors are types of environmental sensors. The outcomes of the sensor-based human activity recognition literature review are presented in the Table 2.3.

Table 2.3: Sensor-Based HAR Studies

Study	Dataset(s)	#Activity	Overlap(%)	WS(s)	CM	Evaluation	Metric	Result
[34]	Collected	12	80	1	k-NN SVM RF HMM GMM	10-Fold	A	0.99 0.95 0.98 0.83 0.72
[41]	Collected	17	50	12.8	SVM	10-Fold	A	0.98
[43]	Collected	22	87.5	40	RF SVM NEV	Holdout	F1	0.75 0.57 0.92
[25]	ExtraSensory	25	NA	20	LR	5-Fold LOO	BA	0.77 (F) 0.78 (L)
[35]	RWHAR	5	50	5	CNN	5-Fold (F) LOO (L)	F1	0.94 (F/M) 0.93 (F/C) 0.75 (L/M) 0.76 (L/C)
[36]	UCI-HAR WISDM	6 6	NA	10 2.56	CNN	10-Fold Holdout	A	0.97 0.93
[38]	UCI-HAR UniMib SHAR WISDM PAMAP2 Opportunity Weakly Labeled	6 17 6 18 18 4	50 50 95 78 50 50	2.56 3 10 5.12 1 40.96	CNN	Holdout	A	0.97 0.76 0.98 0.93 0.93 0.92
[39]	Collected	10	NA	10	CNN	Holdout	F1	0.87
[42]	WISDM	6	50	3.2	CNN	7-Fold	A (A) F1 (F)	0.95 (A) 0.94 (F)
[46]	Ubicomp08(U) Opportunity(O)	34(US) 4(UC) 4(OS) 5(OC)	NA	20 (US) 300 (UC) 0.5 (OS) 10 (OC)	CNN LSTM	LODO Holdout	F1	0.90 (US) 0.90 (UC) 0.92(OS) 0.83(OC)
[40]	UCI-HAR Opportunity UniMib SHAR PAMAP2 WISDM	6 17 17 18 6	50 8stp 2stp 50 20stp	2.56 64rpw 151rpw 2.56 200rpw	CNN	Holdout	A(A) F1(F)	0.96 (A) 0.80 (F) 0.78 (A) 0.92 (A) 0.98 (A)

Table 2.4: Sensor-Based HAR Studies (continued)

Study	Dataset(s)	#Activity	Overlap(%)	WS(s)	CM	Evaluation	Metric	Result
[44]	Opportunity	18	NA	NA	I-NN	Holdout	F1	0.95
	PAMAP2	18	0	5.12				0.94
	UCI-HAR	6	50	2.56				0.95
[37]	MHealth	12	50	2.56	DBN	Holdout	A	0.97
[45]	Collected	6	50	2.56	SLFN(S)	Holdout	A	0.96 (S)
					LSTM(L)			0.97 (L)

*WS:Window Size in seconds, A: Accuracy, BA: Balanced Accuracy, NA:Not Applicable, stp:sliding step length, rpw: row per window

Researchers who introduced ExtraSensory dataset [25] perform HAR task with selected 25 behavioral contexts out of 51 contexts. As cross validation, 5-fold (48 user as train, 12 user as test) and LOO validation methods are selected. They apply Logistic Regression (LR). Paper also emphasis on using balanced accuracy (BA) or F1 score instead of Accuracy (A) due to highly imbalanced structure of activity distributions. In the work, for 5-fold cross validation, 77% BA score is achieved, while for LOO BA accuracy is slightly improved as 78%. In study [34], researchers perform experiments over the data they collected with three inertial sensor (accelerometer, gyroscope and magnetometer) located in upper/lower body. In data pre-processing part, they extract 11 time domain features and 6 frequency domain features from data. They use 80% overlapping window with 25 samples (1 second) each. They also make comparison between raw sensor measurement performance and feature extracted sensor performance of their models and emphasis that with proper features extracted, accuracy values of all classification models they experimented except GMM are increased. Another study [41] focusing on elderly care, where collecting data from both wearable and ambient sensors. In the paper, they investigate performance effect of feature selection methods and compare results. Furthermore, an ambient based hierarchy is represented in this study as another test area. They group activities based on room and conduct experiments with and without feature selection models. Best accuracy score with SVM model is achieved with hierarchical approach with feature selected model CMIM as 98% 10-fold validation.

Deep learning algorithms become dominant architectures recently for HAR tasks. Researcher in [35] represents a CNN model with 3 convolution layers and one fully connected layer. As sensor input, even though the public dataset used in this study has many different sensor located in different parts of the body, only accelerometer from sensor located at waist is used. From raw data, an accelerometer vector is created by substracting activity vector caused by gravity from original vector, to present only human movement. Monochromed Horizontal-Vertical images are created with 50% overlapping window rate. In addition, colored HV images are created by adding time dependency. With 5-fold cross validation, F1 score results obtained are 94% and 93% monochrome and colored HVs respectively. For LOO, F1 score results are 75% and 76%. In [36], a shallow CNN architecture that extracts features automatically to perform real-time operations with best possible time-series duration is represented. Besides classical CNN, researchers combine different pre-processing metrics such as statistical features, data centering, data normalization with CNN to find optimal result. For WISDM, they perform training with 26 user data while remaining rest of the users as test data, which gives them 93% accuracy for 10 second interval. In UCI dataset, they perform 10-fold cross validation for data of 10 users remained and receive 97% A score, for 2.56 second interval. Kernel selection is very important in case of CNN-based studies. Study [38] aims to obtain higher scores than existing studies by examining interchangeable kernel size in the same feature layers. By conducting extensive experiments, they observe the effect of dilation rate and group number changes over accuracy scores with given kernel size. CNN consists of three convolution layers while last two are changed with selective kernel convolution. Table 2.3 lists score of the proposed method for all given datasets. In the paper, they perform comparison with state-of-art methods and receive promising results. Using lightweight CNN architectures for recognition tasks is recently become popular thanks to public datasets. In [40], 5 different experiments are performed with layer-wise approach. There are 3 convolutional layers in this CNN where between each, there are two sub-networks that generates local loss to train weights at each hidden layer. Sub-networks are responsible to calculate similarity matching loss and prediction loss and passes the values to next layer in CNN while effecting weights. Even though holdout validation is used, they divide train and test data differently. Phrase *rpw* represents row per window while *stp* is sliding step length. For UCI-

HAR, UniMib SHAR and WISDM datasets, train and test data rates are 70% and 30%, where with using both of the sub-networks, A scores are 96%, 78% and 98% respectively. By same approach, experiment with PAMAP2 data is divided as 80% as train, rest as test, while achieving A score 92%. Opportunity dataset is highly imbalanced, therefore, score is reported in F1 metric as 80% where, train-test division is done user-based. Traditional end-to-end CNN model approaches circumvent the drawbacks of manual feature extraction. In case there is less train data of a complex activity, recognition performance might be very low. Study [42] tries to tackle this issue by representing a two-stage CNN model in which one part is responsible for only deciding activities *ascending stairs* and *descending stairs* while other does classification for the rest. Study also does data augmentation which respect to duration of action. In data augmentation part, in order to represent short step and long step points same, they perform linear interpolation. Overall accuracy and F1 scores are reported as 95% and 94% respectively for data augmented two-stage CNN. Inception NN structure, which is used in [44], consists of blocks that processes input from previous layer with multiple kernels and concatenate their results. Blocks generally use small sized filters to avoid overfitting for information that locally distributed. In the study, with three are three represented kernel sizes in the model. Holdout validation is applied into datasets that have various overlap ratios and window length. F1 scores over three datasets are reported 95%(Opportunity), 94%(PAMAP2) and 95%(UCI-HAR). Activities could be categorized based on required duration or multiple-single action requirement. In [46], activity categorization is applied based on their complexity such as defining *Standing* activity as simple and *Cooking* as complex. After new data labelling approach, first, data of complex activities are fed into CNN model for feature extraction. Then extracted features are delivered into a softmax layer so that LSTM model performs simple activity recognition from complex ones. In Table 2.3, letters U and O distinguish datasets while letters S and C indicates simple and complex. Ubicomp08 dataset is validated with LODO since it is collected from one person. Opportunity dataset is validated with holdout. Different categories of activities are tested with different window lengths due to nature of activity. Research [37] proposes a DBN framework. DBN-based architecture composed of various number of different units, while they conduct experiments to figure out effect of number of hidden units at different layers of their system. Even though they mentioned using

sliding window technique with 50% overlapping rate with 2.56 second. Study [45] emphasis on benefits of feature engineering over raw sensor data. They extracted time and frequency domain features and use them while conducting performance test with their Single Layer FeedForward Network (SLFN) and LSTM models. LSTM model receive distilled knowledge from SLFN model result at the softmax layer to boost deep model performance by taking benefit from shallow ML methods' performance over features. To use holdout validation method, dataset is divided into two parts, 70% as train and 30% as test data. Accuracy results are reported as 96% for SFLN model and 97% for proposed LSTM model. There are few studies using audio data to perform HAR. In [39], audio data is collected from open source online platforms, where each 9 minutes of record is divided into train for 6 minutes and test for 3 minutes. Melspectrograms are extracted from raw data with 200 mel-frequency bin and 500 frames to conduct experiments. Overall F1 score is 87% while activity *Talking* has highest recognition score, 93%. While [39] uses DL, conventional ML algorithms are used to evaluate performance over human activity recognition based on audio signals. In [43], researchers compare their represented method Non-Markovian Ensemble Voting (NEV) with classical methods, RF and SVM. Each activity stream is learned and tested from generated bag-of-sounds, which is similar to a dictionary showing places of frame starting points and center per each activity audio. Similar to other studies, MFCC features are generated from raw signal for each activity with 87.5% overlapping window rate. F-score average results are reported as 75%, 57%, for classical methods RF and SVM respectively while for their novel method NEV outperforms with 92% F-score.

There are interesting researches in the Table 2.3 that have reported promising results with offered solutions. Studies such as [34], [38], [40] and [41] have achieved very high success results, such as 98% by using different inertial sensors as input. Even though in other studies include less activities, number of activities predicted is 34 in [46], which indicates much more complexity. Still reported F1 score is 90%. There are a few approaches that use audio sensor data to accomplish activity recognition task like [39] which tries to distinct 22 different activities based on sounds recorded. Their novel method NEV reaches 92% F1 score. Sensor-based HAR was used in this thesis since it is an intriguing subject to perform activity recognition with data

gathered from inertial and auditory sensors.

2.4.3.2 Vision Based HAR Studies

Human activity, behavior or mimic-gesture recognition with the help of video recordings' analysis has been extensively researched [62], [63] [47], the area of interest of vision-based HAR. The primary rationale is to use highly informative visual data acquired by cameras to perform high-level interpretation, in order to assist human operators with highest possible efficacy, or, even replace human monitoring with autonomous machines. The approach has a significant impact on security and surveillance [64] [65], healthcare and interactive applications [66]. Table 2.5 lists some of the existing studies for vision-based human activity learning tasks.

In [59], researchers introduced three leveled CNN classifiers for RGB, depth and skeletal data to categorize activity based on motion history images, depth motion maps and skeleton sequence images respectively. For 5 CNN, they use well known VGG-F predefined model. They conduct experiments with 5-fold cross validation and report 95% accuracy for UTD-MHAD dataset, 96% accuracy for SBU-KI dataset which could be interpreted as state of art and 93% accuracy score for CAD-60. Researchers present an framework consists of multiple RNN-trees for skeleton based fine-grained activity recognition in [60]. To begin, they concatenate and process small-scale existing datasets to construct their own dataset including 140 activities. They conduct tests over large scale data which is divided into sets as 60% training, 20% validation and 20% for test. They offer multiple tree structures, most successful model gives 89% accuracy. Also they conduct test to predict model performance over NTU RGB+D dataset. Validation method is not specified, however, most successful result is reported as 83% in this dataset with unidirectional RNN. Study [61] performs good results noisy data. In that study, an three layered SAE with denoising tensor is introducted which manages temporal and spatial corruption of given data. DTAEs cope with different corruption ratios individually when it comes to temporal corruption. Data corruption rates tested are 20%, 40% and 60% respectively. Best results are obtained with 20% noisy data rate, which results 86% accuracy. Framework introduced in [68] is tested within five distinct datasets with distinct activities each.

Table 2.5: Vision-Based HAR Studies

Study	Dataset(s)	#Activity	Classification Model	Evaluation	Metric	Result
[59]	UTD-MHAD(U) SBU-KI(S) CAD-60(C)	27 (U) 8 (S) 12 (C)	CNN	5-Fold	A	0.95(U) 0.96(S) 0.93(C)
[60]	Collected(C) NTU RGB+D(N)	140 (C) 60 (N)	RNN	Holdout (C) NA (N)	A	0.89(C) 0.83(N)
[61]	MSR Action Pairs	12	AE	NA	A	0.86
[68]	UCF101(U) HMDB51(H) Hollywood2(HW) UCF50(UC) YouTube(Y)	101 (U) 51 (H) 12 (HW) 50 (UC) 11 (Y)	CNN & LSTM	Holdout	A	0.94(U) 0.72(H) 0.69(HW) 0.95(UC) 0.96(Y)
[69]	MSR Action3D (A) MSR Daily Activity3D (D)	20 (A) 16 (D)	HMM	LOSO	A	0.93(A) 0.94(D)
[70]	G3D (G) SYSU 3D HOI (S) UTD-MHAD (U) MSR Action3D (A) MSR Daily Activity3D (D)	20 (G) 12 (S) 27 (U) 20 (A) 16 (D)	CNN	LNSO	A	0.96G) 0.95(S) 0.89(U) 1.00(A) 0.97(D)
[71]	MSR Action3D (M) Berkeley MHAD (B) HDM05 (H)	20 (M) 11 (B) 65 (H)	RNN	LNSO	A	0.95 (M) 1.00 (B) 0.97 (H)

*A: Accuracy, BA: Balanced Accuracy, NA: Not Applicable

They concatenate their approach into three phase, preprocessing, feature extraction with CNN and action sequence learning with LSTM. In preprocessing phase, only frames with relavant information are captured from continuous stream. Next, they generate feature maps from frames captured using CNN. In final phase, multilayer LSTM is constructed for sequence learning and predicting the activity based on sequence. Model evaluation is done with dividing dataset into 60% as train, 20% as test and remaining for validation. Proposed methodology outperforms state of art methods in four of the datasets while the reported accuracy result 69% using Hollywood2 does not, based on overall results, even for some classes, there are very high accuracy scores are obtained. In another vision-based work [69], researchers use RGBD data to

extract features from human pose and remove information coming from background. RGB-D images are extracted and used to generate skeletal joint features from captured human pose. Two datasets with distinct number of classes are used with LOSO cross-validation method. HMM model they represented perform in accuracy score unit is 93% for MSR Action3D dataset and 94% with MSR Daily Activity3D dataset. There are multiple processing methods in vision-based HAR such as skeleton data, depth data or RGB. Research [70] presents a methodology for RGB-D video-based action recognition that uses bidirectional rank pooling, using depth primarily. Dynamic depth images are constructed in three level, body, parts and joints respectively. All of those create spatially structured dynamic depth images. Later, bidirectional rank pooling based on depth map sequence is applied with CNN approach to each of those levels separately. As late fusion technique, model achieves higher results with multiply fusion. Table 2.5 shows accuracy scores obtained with proposed model for distinct vision-based HAR datasets. Even though train-test division rates are changed dataset to dataset, generally N user is divided as test while others are used to train model. In case of skeletal joint vision-based study, [71] could be an example. In the study, skeletal joints information of users are hierarchically grouped based on different parts of the body such as leg joints, arm joints etc. An bidirectional hierarchical RNN model is adapted to perform classification task based on classes mentioned in Table 2.5, even though original database might include much more classes. For MSR Action3D dataset, even numbered subjects are selected for testing while odd numbered subjects are for training the model. Overall accuracy score is reported as 95 %. For Berkeley MHAD, data of 7 subjects are used as train data while rest of them is saved for test. Performance of the model in here is reported as 100% with this configuration. Lastly, model performs with 97% for HDM05 dataset, which includes various number of classes.

There are many challenges of vision-based solutions of HAR. Challenges in human activity recognition using multimedia stems not only from the complexity of portraying the motion of body components, but also from a range of other genuine issues such camera motion, a dynamic background, and inclement weather [73] which could effect overall system performance in a negative manner. Furthermore, because human movements are complicated and variable depending on conditions like age,

body-proportion or personal habit, even the tiniest movement could convey context [74], which cause a learning model to mispredict behavior. Although this is an exciting subject, for the reasons stated, a vision-based technique id not chosen for activity recognition in this study.

2.5 Audio Processing

Mel Frequency Cepstrum Cooefficients (MFCC)s are derived from raw audio signal with respect to different transforms and operations. First, continuous signal in waveform is divided into windowing frames, then Discrete Fourier Transform(DFT) is applied to all windowing frames, basically moving the audio signal from time domain to frequency domain. Power Spectrums or on other name Amplitude Spectrums, that show magnitude of the signal changing in frequency are obtained from that operation. They represent perceptually-informed amplitude in frequency. Frequency expression in Hz unit are transformable to frequency expressions in *mel* unit. To convert frequency to mel scale, first, depending on the problem, number of mel bands are selected. Regarding to band range, constructed mel filter banks (generally 32-64) are applied to PSs, as matrix multiplication to receive Mel Power Spectrogram. Melspectrogram expresses frequency(mel bands) change over time. With applying mel-scaling, sound is basically represented in a linear form. Logartihm operation is applied into PS to uniform Log Power Spectrum (Log-PS), which is a continuous frequency with harmonic components that are periodic. Result of that operation is Log Mel Spectrograms. Then, Discrete Cosine Transform (which is simplified version of DFT) is applied into to get the matrix of real-valued cepstral coefficients (traditionally 12-13 coefficients), MFCCs. Figure 2.4 illustrates process to convert waveform signal into first Melspectrogram, then Log Melspectrogram and finally MFCC.

In this study, both Melspectrograms and MFCCs are used to learn from audio signal. Melspectrograms are chosen often to work with DL algorithms while MFCCs are preferred by ML algorithms generally. This work summarizes the performance of both to overcome problem description as well as Log Melspectrogram.

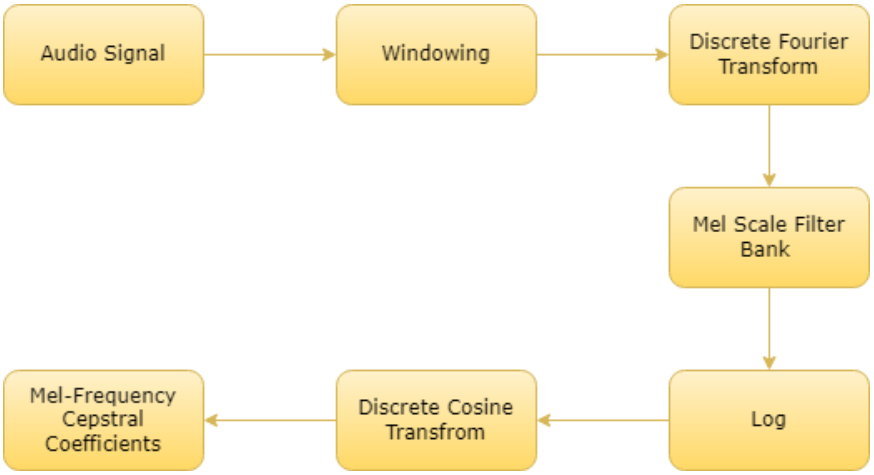


Figure 2.4: Audio Signal Processing steps

CHAPTER 3

PROPOSED FRAMEWORK FOR HUMAN ACTIVITY RECOGNITION

The processes for recognizing human activity using embedded smartphone and smartphone sensors are described in depth in this chapter. First, the framework that depicts overall architecture is introduced. Input generation mechanism that proposed deep learning model would utilize is explained in detail, including preprocessing over data to define clusters and processing technique that translates sensory input for both inertial sensors and audio features. Lastly, to perform recognition task with lesser computational effort, a lightweight deep neural network model is used that performs classification task in which performance is improved by hyperparameter tuning.

3.1 The Framework for Hierarchical Human Activity Recognition

The proposed pipeline in this thesis consists of 5 main steps: (1) data acquisition and feature extraction & selection, (2) data preprocessing and normalization, (3) color coded or/and audio image production, (4) sensor fusion, (5) supervised classification based on hierarchy depth level. Figure 3.1 demonstrates overall framework.

Real-time data collected from smartphone and smartwatch are processed to remove noise at first. After that, features from a variety of sensor modalities are retrieved. In this thesis, only *mean* features of each axis in sensors and *mean* features of MFCC are selected in future selection phase. MSPCs are transformed from MFCCs. Color coded images are produced from the four inertial sensors, where each row represents a single sensor data. Images are also produced utilizing the grayscale to RGB technique from both of the audio features, MFCC and MSPC. The lightweight model learns patterns from each sensor images or fused images of audio and inertial sensors which are

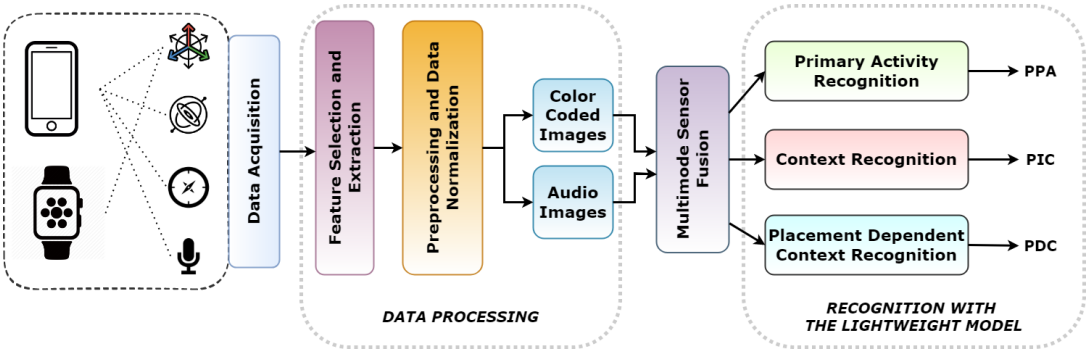


Figure 3.1: Proposed framework for hierarchical human activity recognition

formed in the preceding step. This study introduces a three-tiered hierarchy, primary physical activity at level H1, activity and context pairings at level H2 and finally smartphone placement dependent activity and context pairings at level H3. Thanks to this representation, multi-labeled activity recognition is achieved by using a model adapted to multi-class classification.

3.2 Input Generation For Human Activity Recognition Task

HAR aims to model daily life of people, finding out body motion and user behavior. Models could interpret common movements and behaviors via diverse sensory information. In this section, processing and analysis of data for gathered from four inertial sensors and their fusion with audio sensor is explained. CNN architectures have shown to be extremely successful, particularly in the field of computer vision. Based on their width-height-depth infrastructure, they have shown great performance over visual data. Therefore, activity images with respect to their sequence are constructed from sensory information to take advantage of the architecture. Depthwise construction leads dividing data into channels, resulting into smaller data expressions and lighter model selection that operates faster.

3.2.1 Data Preprocessing

The dataset includes 60 csv files, representing activities of daily lives of 60 users for each timestamp and feature values extracted with feature-engineering techniques. It

is a highly imbalanced dataset since daily or overall contribution of users differ from one to another. Even for single user, since data is not collected in a smartly controlled environment with dedicated sensor measurement units, but it is collected in their natural habitat with naturally used devices, smartphones and smartwatches, distribution of data between activities, contexts are extremely imbalanced. To overcome this issue, in addition to the precautions taken during training, the window width is chosen as 20 point sample (20 row per window) in order to increase the number of images to be obtained from activities containing very little data while generating data. Furthermore, overlap ratio of windowing is chosen as 50% based upon the information in the literature.

Extrasensory includes 51 activity labels. However not all of them indicates physical activities, some of them indicates behavioral contexts, placement of the smartphone or current location the user. Therefore, in this work, original activity labels in dataset are divided into sub-categories and hierarchical approach based on the categories is considered.

Feature data is clustered based on each activity per each user. For hierarchically defined activities, points stamped with labels which exist in each step of the hierarchy are taken from the data. Figure 3.2 shows data clustering flow, for level one user-defined label. At first, each user data segmented based on activity (or contexts depending on hierarchy level) are collected. Later, data is clustered based on sensor information. Each bucket represents that user's sensory data S for issued activity A. This process is iterated for each activity, for each individual. After clustering, windowing operation is performed to construct images.

In this study, besides the inertial sensors, the information gathered from audio signal recorded by smartphone is used to boost learning performance of the model, particularly on specific contexts such as "Talking" or "Watching TV". Raw data from audio signal is not available in the dataset due to privacy reasons. Instead, Mel-frequency Cepstral Coefficients (MFCC), which is numerized representation of sound power spectrum are listed for each timestamp. In the dataset, 26 features are extracted from original audio gathered with smartphone. 13 of them indicates MFCC channels' mean values measured within 400 frame average values and remaining 13 of them represent

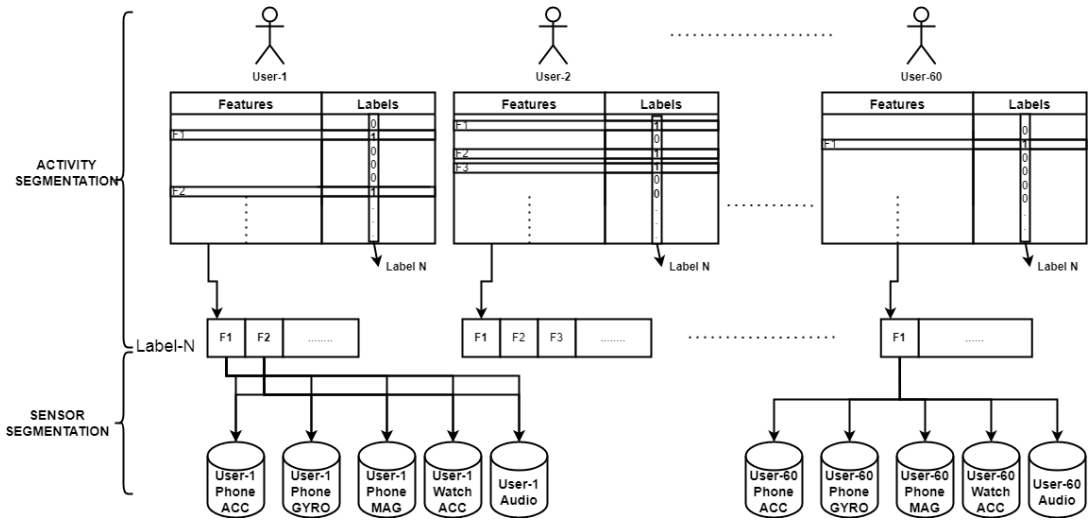


Figure 3.2: Data Clustering Flow

standard deviation values, per each timestamp.

Derivation of MFCCs from audio signals have been explained in section 2.5. Same section also introduces Mel Spectrogram, previous step just before computing MFCC. After applying DCT to Mel Spectrogram, resulted MFCCs are in linear form, which make them vulnerable to be absorbed by neural networks. Therefore, MFCCs are comparably beneficial with linear models, while Mel Spectrograms give better classification performance in DNNs since they are able to learn from more complex structures easily. Both MFCC and Mel Spectrogram representations are used. To receive Mel Spectrograms from MFCCs which are among the features in the dataset, steps mentioned in 2.5 are applied in a reverse manner, with librosa [33] open source library is used. While converting from MFCC to Mel Spectrogram, $nmels$ parameter is given 32, considering existing studies.

3.2.2 Image Generation

Understanding pattern of action or context is a problem to tackle when designing a system. In this study, activities are converted into images to visualize changing sensory information pattern. The aim to represent daily activities or contexts of humans as image is to create motion history in a visualized form which CNN model could utilize the input during feature extraction. Each sensor data has a hardware-dependent

distribution range. Besides, in terms of human activity recognition, a sensor could give values in various distribution ranges for each different activity. Activities follow a general pattern, therefore, sensory information changes are confined for actions require less effort. However, for high effort required actions, sensor measurements indicate a lot of fluctuation. Relying on this approach and considering numerous activities in the data set, first, standard deviation (formula 3.1) value of each sensor measurement's channel is calculated. Then each value in the data is re-calculated with equation 3.2 to normalize values into unit based variance, where i represent element and j represents channel of a sensor. Note that $\overline{\mathrm{x}}$ indicates mean of x. This step of regularization operated user-based and separately per each sensor, because the sensors in smart phones and smart watches that each user uses daily might have various distribution ranges due to varied hardware and software installed. At the same time, this regularization is carried out before activity-based categorization, per each person.

$$\mathbf{s(x)} = \sqrt{\frac{\mathbf{1}}{\mathbf{N-1}} \cdot \sum_{\mathbf{i=1}}^{\mathbf{N}} (\mathbf{x_i} - \overline{\mathbf{x}})^{\mathbf{2}}} \tag{3.1}$$

$$\mathbf{x_{ij}} = \frac{\mathbf{x_{ij}} - \overline{\mathbf{x_j}}}{\mathbf{s(x_j)}} \tag{3.2}$$

In image representation, each pixel has a numerical value and generally they are represented by 8-bits (one-byte), which indicates [0, 255] range. After data regularization explained above has been operated, in order to express those numerical values as pixels of images, another normalization method is applied that maps sensory information with respect to standard deviation proportionally to pixel rate. Mapping values are determined independently for each individual and each sensor data information channel in order to keep natural behavior per person.

In this study, three different sensory image types are created, one for inertial sensors and, two for audio sensor features, MFCC and MSPC. Each one have different size due to feature characteristics. Despite the common functions used in calculating the pixel values, while converting the inertial and audio sensors into images, there are differences in the RGB sequence due to channel numbers. Sensory fusion images are constructed by concatenating inertial and audio feature images.

3.2.2.1 Inertial Sensor Images

Inertial sensors are combination of smartphone sensors (accelerometer, gyroscope, magnetometer) and smartwatch accelerometer. Each of those have 3-channels, X,Y and Z which could be mapped into pixel as RGB channels. Therefore, for inertial sensors, data for X axis is mapped into R, Y axis is mapped into G and Z axis is mapped into B, which led constructing color coded images. This RBG mapping color coding technique is obtained from work of [51]. There are four inertial sensors used in this study, represented by a row in resulting image. Figure 3.3 shows color coded image of activity "Lying Down" at hierarchy level one, each row indicates a sensor information. Color coded images generated based on hierarchy level, figure 3.4 shows an example.

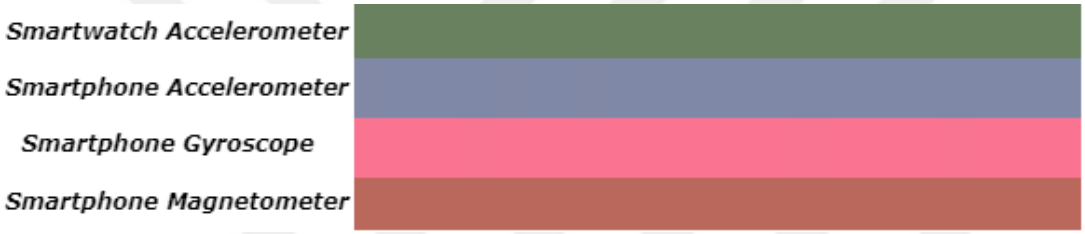
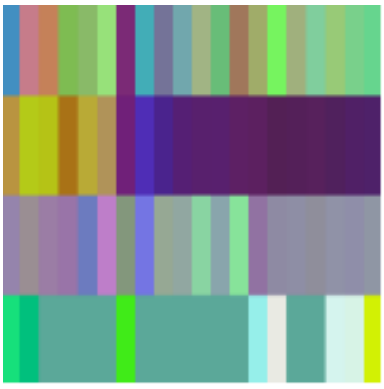
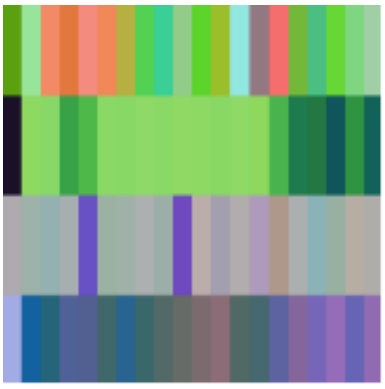


Figure 3.3: Inertial sensory fusion image of Lying Down Activity with Units



(a) H2 Inertial image, Activity: Walking, Context: Shopping, Placement: NA

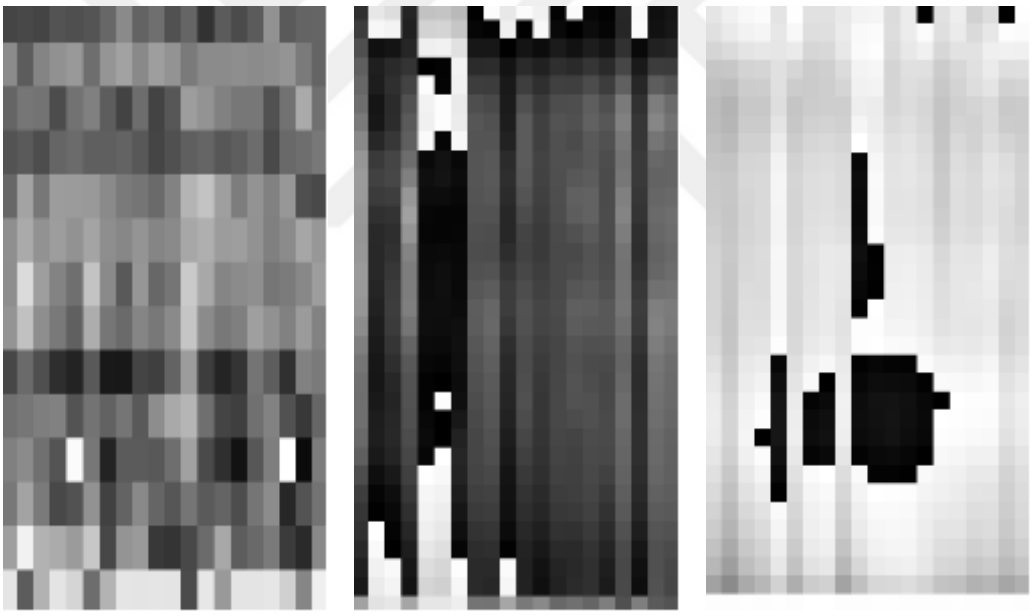


(b) H3 Inertial image, Activity: Standing, Context: Outside, Placement: PIP

Figure 3.4: Inertial images of sample classes with hierarchy

3.2.2.2 Audio Images

The method used for in inertial sensors is not applicable for audio sensor, since MFCC sequences are composed of 13 channels while MSPCs and LMPSCs have 32 channels, and they both have additional feature, maximum value. They could be represented as grayscale images by themselves with one-dimension. However, for training and testing both of inertial sensor images and audio sensor images together to keep overall architecture simple, they are converted to RGB from grayscale, by assigning each value in the channel for all R, G and B channels of the resulting MFCC, Mel Spectrogram or Log Mel Spectrogram image. Therefore, to a novel representation of auido images is introduced in this thesis. This three channelled images are incorporated to color coded images. Figure 3.5 illustrates some examples of MFCC, MSPC and LMPSC based images.



(a) H2 MFCC image, Activity: Walking, Context: Talking, Placement: NA

(b) H3 MSPC image, Activity: Sitting, Context: In a Meeting, P: Phone on Table

(c) H1 LMSPC image, Activity: Bicycling, Context: NA, Placement: NA

Figure 3.5: Audio Feature Images

3.3 Convolutional Neural Network Model

To operate classification based on activity, images representing sensor fusion for each activity are constructred after preprocessing and color coding. Mathematical convolution operation enables to extract information based on local neighborhood with respect to kernel weights. CNN algorithms that takes this advantage are exceptionally good at lowering the amount of parameters without sacrificing model quality, considering inputs that have high dimensionality such as images. Architecture used here is inspired by work of [51]. However, to achieve best results out of model, some parts are tested with different values. The task of selecting a collection of ideal hyperparameters for a learning algorithm is known as hyperparameter tuning. A hyperparameter is a value for a parameter that is used to influence the learning process, determined before training. In this work, to find optimal values for each configurable layer in the network, hyperparameter tuning is performed. Other factors, such as node weights are, on the other hand, learned during training.

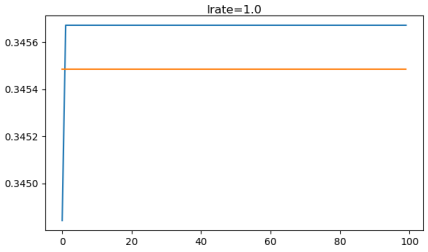
Kernel size is crucial element of mathematical convolution that effects directly the feature extraction process. Small kernel size is beneficial to detect details while bigger kernel size could detect largely distributed features. Also, odd-sized kernels are preferred rather than even ones in order to maintain symmetry of pixels. Without that symmetry, distortions might occur across consecutive layers, therefore, odd kernel size is safer choice, rather than handling any possible corruption. Images consist of different heights due to inertial, MFCC, Melspectrogram and combined images. Therefore, using large kernel might reduce features coming from inertial sensors while expiring for audio-based sensor. As a result, 3x3 kernel size is selected.

Feature map or in another name activation dimensions might vary based on configuration. It's a mapping that correlates to activation of different areas of the image, and it's also known as a feature map, because, it points where a specific type of feature could be located inside the output image with respect to filter. There are no significant studies that indicate the effect of higher feature map sizes increase model performance. In order to keep lightweight of the model and number of training parameters, number of feature map is selected 32 for all convolutional layers.

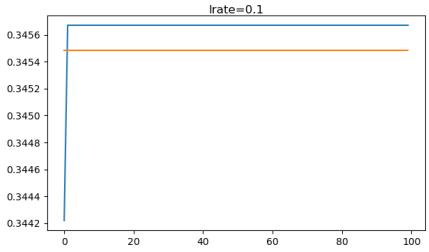
Pooling layer is responsible of dimension reduction over feature maps by eliminating redundant data, based on the selected method. As a result, it has a considerable impact on feature extraction process. There are two main methods, max pooling and average pooling to summarize features in an area of the image. In this work, max pooling method is used. As stated before, each row in the image inertial area, first four columns since there are four inertial sensors used in this study, represents a distinct sensor, while for audio sensors, image height is bigger. In order to avoid losing important features from inertial sensors, pooling layer kernel size is selected as 1x2 to operate dimension reduction row-based.

Fully connected or dense layers do classification operation with data fed to them. In order avoid deciding classification results directly with one fully connected layer which has number of class as output channel, in addition to that layer, additional two fully connected layers are used. First fully connected layer linked to flatten layer has 512 neurons, this layer is followed by another fully connected layer with 256 neurons. Those two use ReLU as activation function. The last layer has neurons equal to number of classes, which might change based on hierarchy or test group. Final dense layer works with softmax activation function.

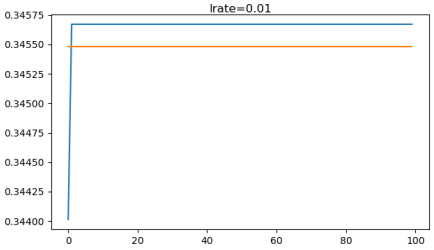
There are many optimizers, but in this work, Sthocastic Gradient Descent (SGD) is selected. Gradients multiplied with specified learning rates are substracted from the weights with this optimizer. In this case, choosing best learning rate for system might increase system accuracy. Given the smaller changes to the weights for each update, smaller learning rates necessitate more training epochs, whereas greater learning rates necessitate fewer training epochs. Therefore, learning rate selection should be considered with number of epochs. On the other hand, momentum is another choice effects performance of SGD. Momentum conceptually states acceleration rate in the training process. Figure 3.6 illustrates the results of experiments performed to select the most robust learning rate.



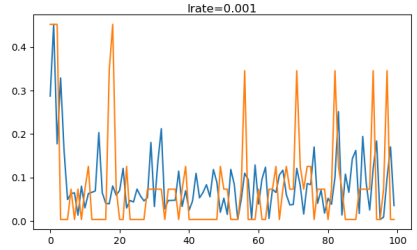
(a) Learning Rate = 1.0



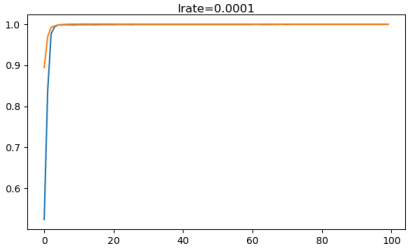
(b) Learning Rate = 0.1



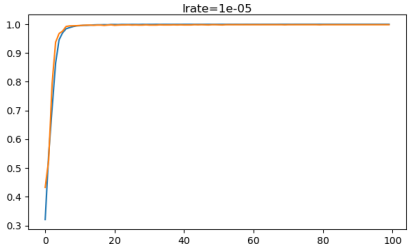
(c) Learning Rate = 0.01



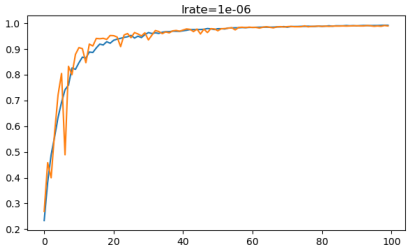
(d) Learning Rate = 0.001



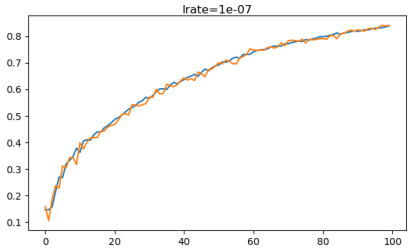
(e) Learning Rate = 0.0001



(f) Learning Rate = 1e-05



(g) Learning Rate = 1e-06



(h) Learning Rate = 1e-07

Figure 3.6: Accuracy vs Epoch Graphs of each Learning Rate

To sum up, the CNN based architecture defined here consists of 4 convolutional layers that each have 32 output channels and 3x3 kernel size. Each convolution layer is followed by a max pooling layer with kernel size 1x2. Activation function is chosen as Rectified Linear Unit (ReLU) for both convolution and max pooling layers. After last pooling layer, a flatten layer is added into the model to reshape dimension of features fed to first fully connected layer, with 512 hidden units. Next fully connected layer has 256 hidden units. Finally, hidden units of the last fully connected layer are equal to number of classes, which changes based on test cases. The activation function of the last fully connected layer is softmax, that performs well with multiple class problem. As optimizer, SGD is used with learning rate 0.0001 and momentum 0.9 while loss function is selected as Categorical Cross-Entropy Loss. To make a robust prediction, to keep validation loss minimum and keep accuracy maximum, an early stop condition is defined that waits until 50th epoch, but mainly stops when the loss quantity being measured after each epoch stops decreasing, in other words, minimum validation loss is encountered during training. Overall structure is in figure 3.7 and details of the model are listed below:

- Number of Convolotional Layers: 4
- Number of Max Pooling Layers: 4
- Number of Dense Layers: 3
- Activation Function: Rectified Linear Unit (ReLU) and Softmax
- Convolution Kernel: 3x3
- Max Pooling Kernel: 1x2
- Number of Hidden Layers per Dense Layer: 512 (Layer 1), 256 (Layer 2)
- Optimizer: Sthocastic Gradient Descent (SGD)
- Learning Rate and Momentum: 0.0001 and 0.9
- Batch Size: 32
- Epochs: 100

- Loss Function: Categorical Cross-Entropy
- Early Stop Condition: Minimum validation loss

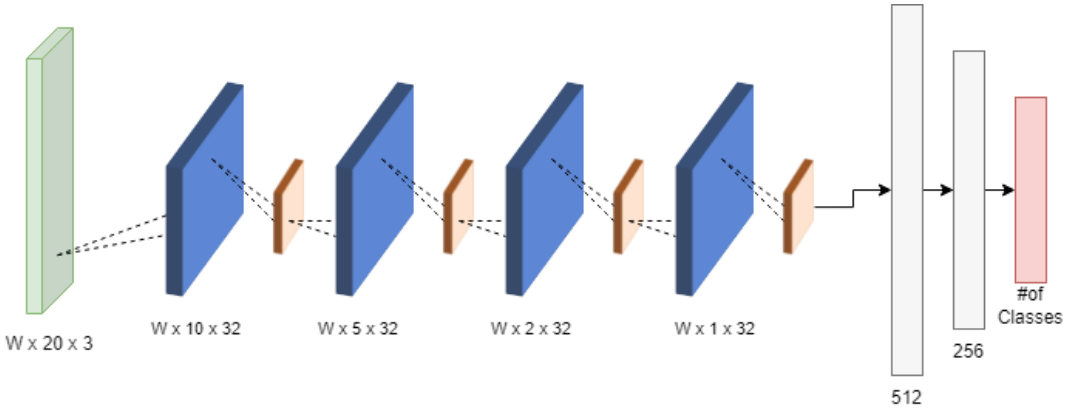


Figure 3.7: CNN structure used in this work

Keras [72] library is used to build model and perform tasks in this study, since it provides simple interfaces with as little user input as feasible.

CHAPTER 4

EXPERIMENTS AND EVALUATION

This chapter covers experiments and evaluation of Human Activity Recognition method represented in this work. Deep Learning model represented in section 3.3 is evaluated. Experiments are conducted in two main ways. The first one includes testing all user defined activities based on their primary physical activity within their hierarchy level without grouping. The second one is testing each activity within its corresponding group based on primary physical activity in its hierarchy level. During model evaluation under this chapter, two cross-validation methods are applied; holdout and 5-fold. For holdout, data is randomly divided into three parts as train, validation and test, per each user-defined label chunk to minimize misleading results due to unbalanced nature of the real-world data. Data division rates are 70% as train, 10% as validation and 20% as test, while conduction experiments for all types. Validation dataset is used during early stopping and controlling whether there are any overfitting in the system or not. This process is repeated per each holdout validation, which is five for each of the testing strategies. For 5-fold cross validation, data is firstly shuffled, then divided into five parts. Per each fold, single part becomes test data and remaining four parts are used as train data. To do so, it is guaranteed that whole generated data is used as train and test purposes. For early stop strategies, test dataset is used to prevent overfitting. Extrasensory dataset represents real-life of subjects. As it is emphasized during the dataset review process, since tests are conducted with an highly imbalanced dataset, balanced accuracy scores or F1 scores are used to evaluate performance. Furthermore, to improve performance of activities with less data, *class weight* approach is applied. Despite the number of samples from each class in the training data, class weights give all classes equal influence on gradient changes. This prevents models from overestimating the class with more samples over lesser ones.

4.1 Hierarchical Evaluation based on Main Group Independency

In this section, primary physical activities (H1), placement independent contexts (H2) and placement dependent contexts (H3) are tested with labels hierarchically at the same level with them without grouping based on prior activity. Position dependency is based on smartphone settlement, in which main sensory measurement units are located. Figure 4.1 represents user-defined labels hierarchically combined with original labels in the dataset, in which tests under this section are conducted.

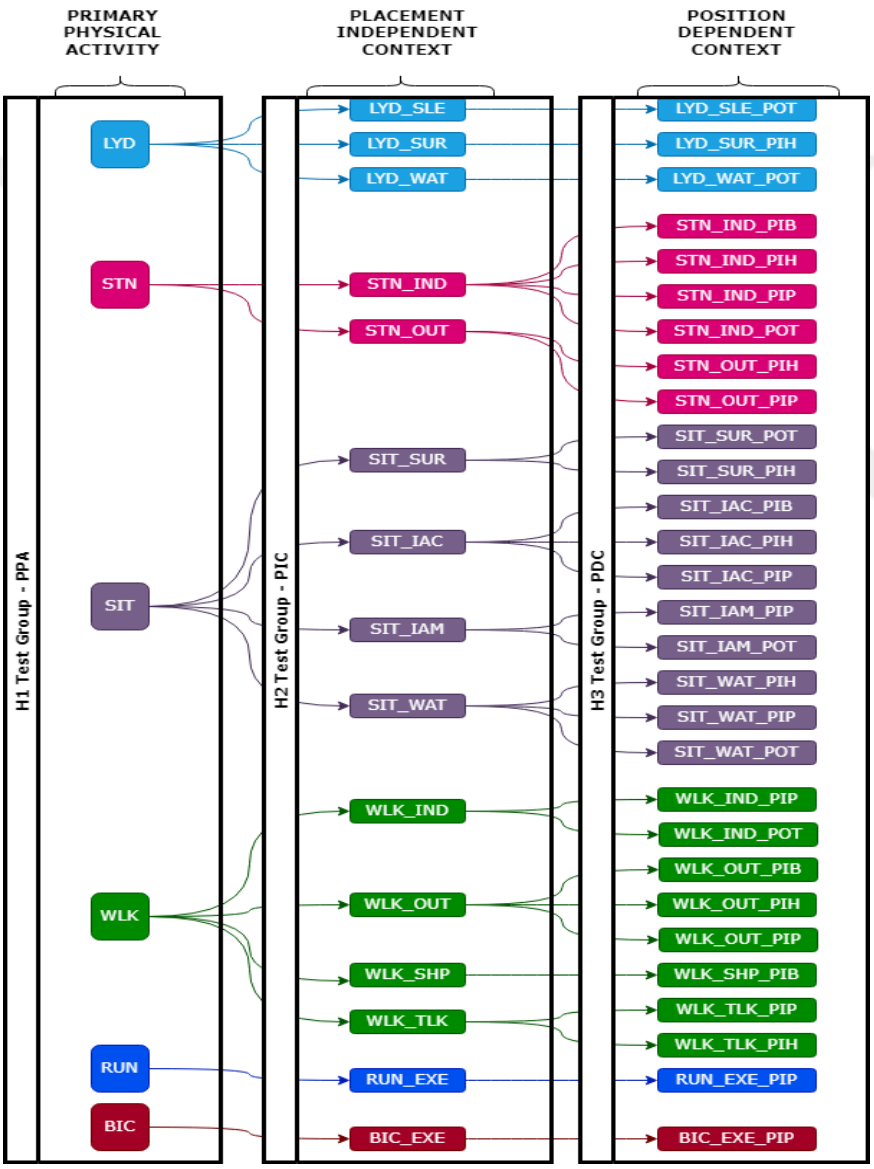


Figure 4.1: Hierarchically User-Defined Activities and Test Groups - Main Group Independent

4.1.1 Holdout Cross Validation Results

The tests are conducted under this subsection with holdout cross validation method where details are explained above. Holdout validation method is applied five times and average results are listed.

In hierarchy level one (H1), recognising operation is performed for six main physical activities with all inertial, audio and their combination data. *H1 Test Group - PPA* clusterization in Figure 4.1 represents activities tested. Table 4.1 lists results of experiments. According to results gathered, audio data itself is very informative to distinct physical activity as well. In case of training with MSPC features only, mean balanced accuracy score of 85% highlights the importance of the impact of sound in recognizing primary physical activities. However, by combining with inertial sensor modalities in color coded image base, overall result is improved as 90 %, which is a very promising data considering different user habits, in case of using MSPC. MFCC and inertial sensor fusion features also present promising results with 88 %. Figure 4.2 demonstrates overall F1 score per primary physical activity. Audio impact could be seen conspicuously from the figure. The precise activity detection results are substantially improved when the audio sensor is paired with the inertial sensors. For example, audio effect boosts up performance to distinct physical activity "Running" among other ones, even though inertial sensors are insufficient. Simultaneously, it has been observed that the audio sensor alone could execute the activity recognition operation with a sufficient accuracy for some activities. Furthermore, it could be seen

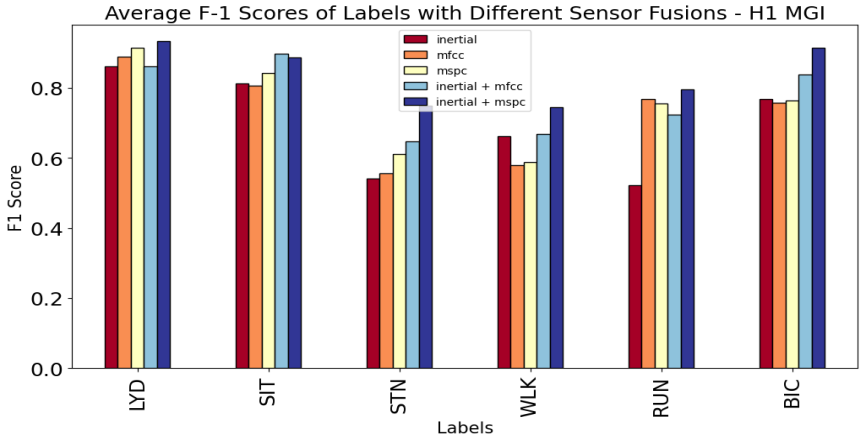


Figure 4.2: PPA Main Group Independent F1 scores per activity - Holdout

that MSPC performance is marginally better than MFCC performance for primary activities, and as a result, combination performance.

Table 4.1: Holdout Results of Primary Physical Activities

Sensor	Holdout ID	ACC	REC	SPE	BACC	PRE	Mac-F1	Mic-F1	W-F1
Inertial	1	0.767	0.692	0.943	0.817	0.680	0.683	0.767	0.768
	2	0.786	0.717	0.946	0.832	0.723	0.717	0.786	0.788
	3	0.769	0.700	0.943	0.822	0.685	0.690	0.769	0.771
	4	0.785	0.710	0.946	0.828	0.722	0.712	0.785	0.786
	5	0.786	0.707	0.946	0.827	0.720	0.711	0.786	0.785
	Mean	0.779	0.705	0.945	0.825	0.706	0.702	0.779	0.780
MFCC	1	0.785	0.709	0.947	0.827	0.741	0.723	0.785	0.781
	2	0.794	0.728	0.949	0.839	0.750	0.738	0.794	0.791
	3	0.782	0.699	0.946	0.822	0.725	0.710	0.782	0.780
	4	0.795	0.718	0.950	0.834	0.750	0.732	0.795	0.793
	5	0.795	0.711	0.950	0.830	0.747	0.727	0.795	0.794
	Mean	0.790	0.713	0.948	0.830	0.743	0.726	0.790	0.788
MSPC	1	0.802	0.726	0.953	0.840	0.738	0.730	0.802	0.803
	2	0.812	0.741	0.956	0.849	0.753	0.745	0.812	0.815
	3	0.827	0.737	0.958	0.847	0.763	0.749	0.827	0.825
	4	0.823	0.746	0.958	0.852	0.782	0.762	0.828	0.831
	5	0.826	0.735	0.957	0.846	0.763	0.748	0.826	0.829
	Mean	0.819	0.737	0.956	0.847	0.760	0.747	0.819	0.820
Inertial & MFCC	1	0.842	0.791	0.961	0.876	0.759	0.774	0.842	0.842
	2	0.839	0.783	0.961	0.872	0.763	0.771	0.839	0.831
	3	0.845	0.805	0.962	0.884	0.772	0.787	0.845	0.846
	4	0.845	0.797	0.962	0.879	0.762	0.777	0.845	0.845
	5	0.840	0.794	0.961	0.878	0.761	0.775	0.840	0.841
	Mean	0.842	0.794	0.961	0.878	0.752	0.776	0.842	0.842
Inertial & MSPC	1	0.876	0.848	0.970	0.909	0.840	0.842	0.876	0.877
	2	0.860	0.823	0.966	0.894	0.816	0.818	0.860	0.859
	3	0.873	0.824	0.969	0.896	0.842	0.831	0.873	0.875
	4	0.869	0.826	0.968	0.897	0.832	0.827	0.869	0.869
	5	0.867	0.808	0.968	0.888	0.848	0.824	0.867	0.868
	Mean	0.869	0.826	0.968	0.897	0.836	0.828	0.869	0.870s

In hierarchy level two (H2), recognising operation is performed for 15 human contexts where each one includes at least one primary physical activities with all inertial, audio and their combination data.

Table 4.2: Holdout Results of Placement Independent Contexts

Sensor	Holdout ID	ACC	REC	SPE	BACC	PRE	Mac-F1	Mic-F1	W-F1
Inertial	1	0.718	0.528	0.977	0.753	0.521	0.520	0.718	0.722
	2	0.713	0.474	0.977	0.726	0.483	0.472	0.713	0.717
	3	0.698	0.495	0.977	0.736	0.486	0.485	0.698	0.703
	4	0.713	0.490	0.977	0.733	0.513	0.495	0.713	0.717
	5	0.710	0.481	0.977	0.729	0.493	0.479	0.710	0.711
	Mean	0.711	0.494	0.977	0.735	0.499	0.490	0.711	0.715
MFCC	1	0.767	0.592	0.982	0.787	0.589	0.588	0.767	0.771
	2	0.748	0.522	0.981	0.751	0.530	0.522	0.748	0.749
	3	0.758	0.532	0.982	0.757	0.509	0.516	0.758	0.757
	4	0.753	0.524	0.981	0.752	0.591	0.527	0.753	0.758
	5	0.767	0.546	0.982	0.764	0.567	0.554	0.767	0.769
	Mean	0.758	0.543	0.981	0.762	0.557	0.541	0.758	0.761
MSPC	1	0.777	0.585	0.982	0.783	0.582	0.580	0.777	0.776
	2	0.776	0.572	0.982	0.777	0.598	0.583	0.776	0.775
	3	0.768	0.532	0.982	0.757	0.594	0.548	0.768	0.769
	4	0.751	0.509	0.980	0.745	0.542	0.521	0.751	0.751
	5	0.768	0.573	0.981	0.777	0.581	0.574	0.768	0.767
	Mean	0.768	0.554	0.982	0.768	0.579	0.561	0.768	0.767
Inertial & MFCC	1	0.802	0.616	0.985	0.800	0.608	0.610	0.802	0.801
	2	0.792	0.585	0.984	0.784	0.606	0.590	0.792	0.793
	3	0.796	0.598	0.984	0.791	0.605	0.600	0.796	0.795
	4	0.805	0.603	0.985	0.794	0.618	0.606	0.805	0.803
	5	0.804	0.600	0.985	0.792	0.627	0.611	0.804	0.804
	Mean	0.800	0.600	0.984	0.792	0.613	0.604	0.800	0.799
Inertial & MPSC	1	0.801	0.560	0.984	0.772	0.617	0.579	0.801	0.802
	2	0.798	0.609	0.983	0.796	0.635	0.615	0.798	0.795
	3	0.793	0.606	0.983	0.795	0.620	0.610	0.793	0.791
	4	0.803	0.619	0.985	0.802	0.637	0.622	0.803	0.802
	5	0.795	0.591	0.984	0.788	0.616	0.601	0.795	0.797
	Mean	0.798	0.597	0.984	0.790	0.625	0.605	0.798	0.797

The labels used are presented in Figure 4.1 within cluster *H2 Test Group - PIC*, indicates hierarchy depth two, 15 phone placement independent contexts. Table 4.2 lists experiment results of each iteration. Compared to previous layer in hierarchy, scores tend to downscale. However, there are multiple context issued here. The importance of audio and inertial sensor fusions is seen when looking at the performance outcomes. Without any audio help, predictability success become 75% model trained only with inertial sensory information. It is observed that among others, "Standing, Outside" context is much more prone to be confused with others. When any kind of audio, whether MFCC or MSPC feature is combined with inertial sensors, however, balanced accuracy score is increased to 80% in best case, 79% average.

Context-based scores are illustrated with Figure 4.3. Similar to the higher levels, fusion of both inertial and audio sensors demonstrate better results compared to others, in general. Especially, for contexts such as "Watching TV" which includes constant background noise or "In a Meeting" includes human interaction, recognition performance is extremely enhanced compared to inertial sensors.

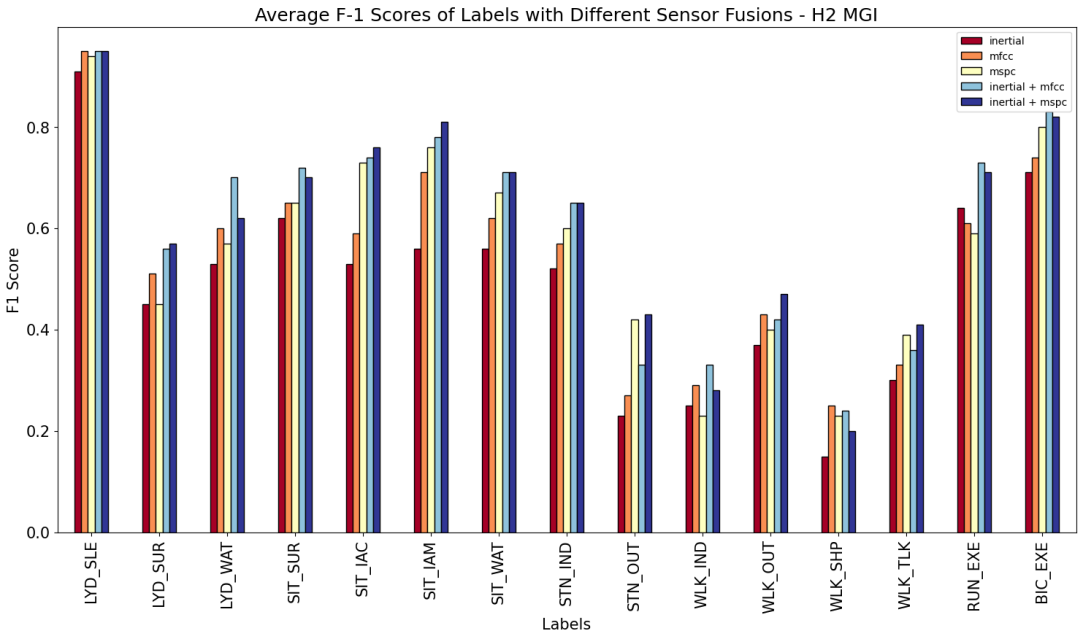


Figure 4.3: PIC Main Group Independent F1 scores per context - Holdout

From the Figure 4.3, it is observed that overall performance of some contexts are very high while others are low. Main reason behind that issue is unbalanced distribution of real-world data both inter-user and intra-user based.

Lastly, user defined labels based on placement dependent contexts in hierarchy level three (H3) are evaluated, results are listed in Table 4.3. Figure 4.1 shows 3rd hierarchy classes listed under cluster *H3 Test Group - PDC*. The system tries to understand activity, context as well as smartphone placement in this case, so, user-defined activity count is increased as 28. Note that, *Running & Exercise & Phone in Pocket* context images could not be constructed due to windowing limitations, therefore, hierarchy level three does not include any context coming from *Running* primary activity. The overall results are slightly lower than the higher level.

At this level, sensor fusion difference become more conspicuous. By using inertial sensors' information only, balanced accuracy score become 72.5%. Overall system performance is boosted up to 77% by synthesizing two modalities, audio and inertial together. In inertial only case, system tend to be confused based on smartphone placement. With the help of audio to inertial sensor information, the model prediction rate is increased for contexts including environmental sounds.

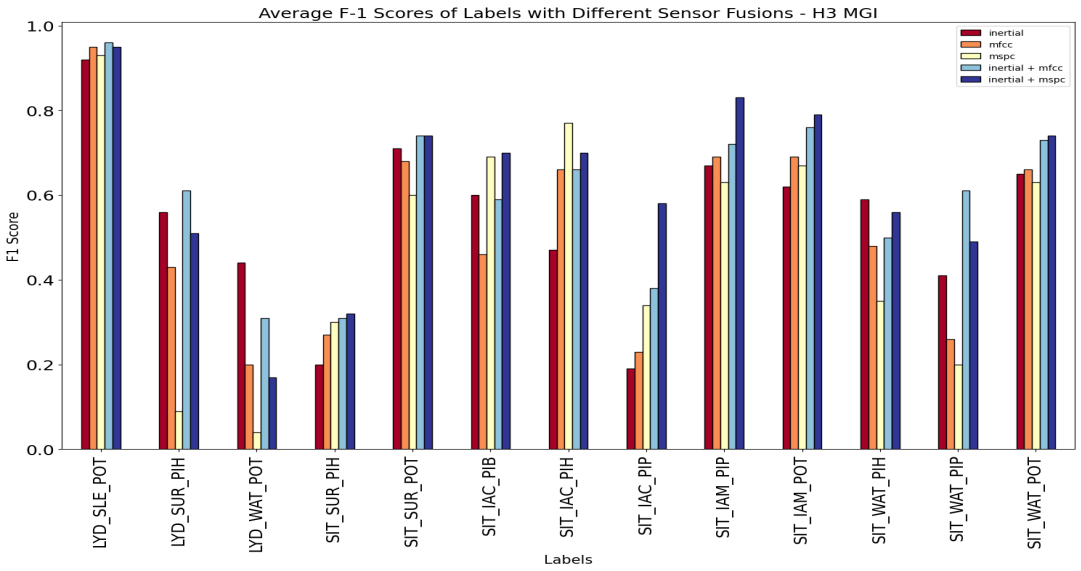
Table 4.3: Holdout Results of Placement Dependent Contexts

Sensor	Holdout ID	ACC	REC	SPE	BACC	PRE	Mac-F1	Mic-F1	W-F1
Inertial	1	0.729	0.478	0.989	0.734	0.471	0.462	0.729	0.738
	2	0.738	0.442	0.989	0.716	0.487	0.444	0.738	0.742
	3	0.752	0.495	0.990	0.742	0.531	0.500	0.752	0.755
	4	0.729	0.425	0.988	0.707	0.437	0.421	0.729	0.731
	5	0.733	0.468	0.989	0.729	0.491	0.466	0.733	0.742
	Mean	0.736	0.462	0.989	0.725	0.483	0.459	0.736	0.742
MFCC	1	0.761	0.438	0.990	0.714	0.564	0.469	0.761	0.767
	2	0.757	0.444	0.990	0.717	0.545	0.457	0.757	0.758
	3	0.764	0.492	0.990	0.741	0.548	0.502	0.764	0.768
	4	0.738	0.452	0.989	0.721	0.469	0.446	0.738	0.737
	5	0.744	0.464	0.989	0.727	0.469	0.458	0.744	0.746
	Mean	0.753	0.458	0.990	0.724	0.519	0.466	0.753	0.755

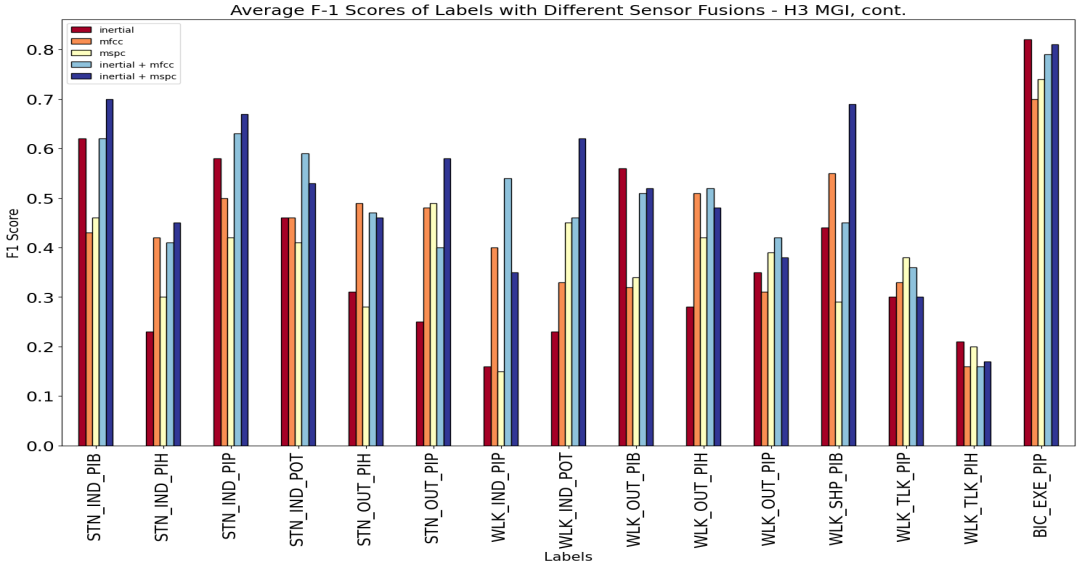
Table 4.4: Holdout Results of Placement Dependent Contexts (continued)

Sensor	Holdout ID	ACC	REC	SPE	BACC	PRE	Mac-F1	Mic-F1	W-F1
MSPC	1	0.751	0.449	0.989	0.719	0.508	0.464	0.751	0.755
	2	0.736	0.425	0.989	0.707	0.474	0.428	0.736	0.738
	3	0.751	0.457	0.990	0.723	0.527	0.469	0.751	0.756
	4	0.689	0.363	0.987	0.675	0.390	0.367	0.689	0.692
	5	0.704	0.399	0.988	0.693	0.455	0.410	0.704	0.704
	Mean	0.726	0.419	0.989	0.704	0.471	0.428	0.726	0.727
Inertial & MFCC	1	0.808	0.549	0.992	0.770	0.609	0.564	0.808	0.810
	2	0.811	0.573	0.992	0.783	0.626	0.585	0.811	0.814
	3	0.798	0.573	0.992	0.782	0.599	0.571	0.798	0.799
	4	0.802	0.546	0.992	0.769	0.575	0.552	0.802	0.807
	5	0.804	0.524	0.992	0.758	0.623	0.546	0.804	0.804
	Mean	0.805	0.553	0.992	0.772	0.606	0.563	0.805	0.807
Inertial & MSPC	1	0.769	0.522	0.990	0.756	0.559	0.527	0.769	0.768
	2	0.804	0.563	0.992	0.777	0.604	0.570	0.804	0.807
	3	0.776	0.530	0.991	0.760	0.659	0.564	0.776	0.778
	4	0.794	0.536	0.992	0.764	0.578	0.540	0.794	0.794
	5	0.815	0.608	0.993	0.800	0.624	0.608	0.815	0.816
	Mean	0.792	0.552	0.991	0.772	0.605	0.562	0.792	0.793

In case of PDC performance evaluation under this group, Figure 4.4 demonstrates placement dependent context performances overall 28 contexts, with F-1 scores. Concluded from the figure, overall performance of the model that utilizes combined modalities is resulted higher than the single modalities, only inertial or only audio. Best fusion combination differ from user-defined activity to activity. For some cases, MSPC and audio combination gives better prediction scores while for some cases, MFCC and audio combination outperforms others. For example, for "Sitting, In a Meeting, Phone on Table" PIC, MSPC and audio fusion gives best results while for "Walking, Indoors, Phone in Pocket" PIC, MFCC and audio fusion resulted best F-1 scores. Sensor fusion enables model to differentiate contexts even smartphone placement is in same configuration, which could model to mispredict current context only using inertial sensors.



(a)



(b)

Figure 4.4: PDC Main Group Independent F1 scores per context-placement - Holdout

4.1.2 5-Fold Cross Validation Results

The tests are conducted under this subsection with 5-fold cross validation method where details are explained above. Results using each parts as test data once are calculated and mean results are listed.

In hierarchy level one, for main group independent testing part, there are six physical activities, could be seen in 4.1. Table 4.5 lists the results for different input types.

Table 4.5: 5-Fold Results of Primary Physical Activities

Sensor	Fold ID	ACC	REC	SPE	BACC	PRE	Mac-F1	Mic-F1	W-F1
Inertial	1	0.774	0.716	0.945	0.830	0.666	0.687	0.774	0.775
	2	0.757	0.695	0.941	0.818	0.653	0.671	0.757	0.759
	3	0.766	0.705	0.943	0.824	0.661	0.680	0.766	0.768
	4	0.768	0.705	0.944	0.824	0.661	0.679	0.768	0.769
	5	0.754	0.690	0.940	0.815	0.649	0.667	0.754	0.755
	Mean	0.764	0.702	0.943	0.823	0.658	0.677	0.764	0.765
MFCC	1	0.794	0.701	0.948	0.824	0.740	0.716	0.794	0.790
	2	0.801	0.717	0.950	0.833	0.748	0.728	0.801	0.798
	3	0.806	0.724	0.951	0.837	0.752	0.734	0.806	0.804
	4	0.786	0.693	0.947	0.820	0.700	0.695	0.786	0.784
	5	0.805	0.731	0.951	0.841	0.754	0.740	0.805	0.804
	Mean	0.798	0.713	0.949	0.831	0.739	0.722	0.798	0.796
MSPC	1	0.816	0.722	0.954	0.838	0.755	0.737	0.816	0.817
	2	0.790	0.745	0.950	0.848	0.738	0.738	0.790	0.793
	3	0.799	0.708	0.950	0.829	0.788	0.741	0.799	0.797
	4	0.783	0.737	0.949	0.843	0.731	0.730	0.783	0.786
	5	0.795	0.749	0.951	0.850	0.745	0.743	0.795	0.798
	Mean	0.797	0.732	0.951	0.842	0.751	0.738	0.797	0.798
LMSPC	1	0.784	0.680	0.947	0.814	0.707	0.693	0.784	0.783
	2	0.775	0.693	0.944	0.819	0.721	0.706	0.775	0.771
	3	0.774	0.687	0.945	0.816	0.746	0.711	0.774	0.774
	4	0.762	0.669	0.941	0.805	0.701	0.683	0.762	0.758
	5	0.779	0.698	0.946	0.822	0.725	0.710	0.779	0.777
	Mean	0.775	0.686	0.945	0.815	0.720	0.700	0.775	0.773
Inertial & MFCC	1	0.882	0.832	0.972	0.902	0.826	0.829	0.882	0.882
	2	0.861	0.819	0.967	0.893	0.814	0.816	0.861	0.862
	3	0.886	0.831	0.973	0.902	0.855	0.841	0.886	0.887
	4	0.886	0.831	0.973	0.902	0.855	0.841	0.854	0.854
	5	0.886	0.842	0.973	0.908	0.856	0.848	0.886	0.887
	Mean	0.874	0.826	0.970	0.898	0.832	0.828	0.874	0.874

Table 4.6: 5-Fold Results of Primary Physical Activities (continued)

Sensor	Fold ID	ACC	REC	SPE	BACC	PRE	Mac-F1	Mic-F1	W-F1
Inertial & MSPC	1	0.884	0.865	0.972	0.919	0.829	0.843	0.884	0.885
	2	0.870	0.846	0.969	0.907	0.813	0.827	0.870	0.870
	3	0.881	0.853	0.972	0.912	0.823	0.835	0.881	0.882
	4	0.875	0.846	0.970	0.908	0.816	0.828	0.875	0.875
	5	0.885	0.855	0.972	0.914	0.827	0.838	0.885	0.886
	Mean	0.879	0.853	0.971	0.912	0.822	0.834	0.879	0.880
Inertial & LMSPC	1	0.873	0.814	0.969	0.892	0.828	0.821	0.873	0.873
	2	0.863	0.792	0.966	0.879	0.833	0.811	0.863	0.862
	3	0.866	0.801	0.967	0.884	0.838	0.818	0.866	0.865
	4	0.873	0.809	0.969	0.889	0.827	0.817	0.873	0.872
	5	0.877	0.811	0.970	0.890	0.830	0.820	0.877	0.876
	Mean	0.871	0.806	0.968	0.887	0.831	0.817	0.871	0.870

The model performs with overall 82% with using inertial sensors only, as could be seen from the results. However, single sound sensors indicate better results than inertial in case of MFCC and MSPC features, while LMSPC is slightly lower. Among three audio sensor types, MSPC is the most informative one with prediction score 84%, where 83% is achieved with MFCC and 82% with LMSPC. The sensor fusion effect is noticeable in the results. The average balanced accuracy score is measured as 91% for inertial and MSPC sensor fusions, while other fusion types give very accurate results as well. Figure 4.5 illustrates average F1 scores measured per each fold per each classes. The figure clearly demonstrates the audio impact. When the auditory sensor is used in conjunction with the inertial sensors, the accuracy of activity detection is greatly improved. For instance, in the case of the "Bicycling" activity, the bar graph shows that the auditory component MSPC aids the model in identifying that activity. Additionally, it is discovered that for some activities, the audio sensor alone could perform the activity recognition operation with adequate accuracy, especially feature MSPC of audio in this case.

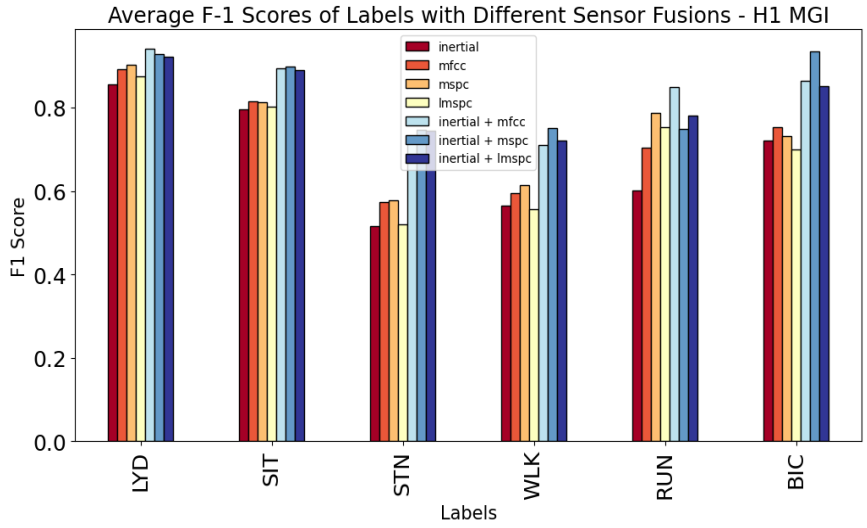


Figure 4.5: PPA Main Group Independent F1 scores per activity - 5-Fold

In hierarchy level two (H2), multi-class classification is operated among 15 distinct activity-context pair classes represented in Figure 4.1, which indicates Placement Independent Contexts (*PIC*). The test are conducted by applying 5-fold cross-validation. Table 4.7 includes all fold results in different metrics and average. When only inertial sensors' information is employed, the model's prediction score is measured at 75%. The auditory features imply higher scores than the inertial sensor alone. MSPC appears to be the most informative of the three audio elements when compared to the others, with balanced accuracy score 79%. The enhancing effect of audio and inertial sensor fusion may be deduced from the findings. The model performance is highest when inertial and MFCC fusion is used, with an nearly 81% balanced accuracy score. Inertial-MSPC and inertial-LMSPC scores are both slightly lower than this number, both of them achieve 80% balanced accuracy score. These results show that when auditory sensors are combined with inertial sensors, they produce excellent results not just in recognizing physical activity but also in identifying contexts. Simultaneously, results are acquired demonstrating how effective audio sensors are at recognizing context, regardless from the most suitable type.

Table 4.7: 5-Fold Results of Placement Independent Contexts

Sensor	Fold ID	ACC	REC	SPE	BACC	PRE	Mac-F1	Mic-F1	W-F1
Inertial	1	0.726	0.543	0.979	0.761	0.520	0.522	0.726	0.730
	2	0.722	0.515	0.978	0.747	0.506	0.504	0.722	0.726
	3	0.712	0.507	0.978	0.742	0.505	0.499	0.712	0.719
	4	0.732	0.555	0.979	0.767	0.527	0.538	0.732	0.736
	5	0.732	0.510	0.979	0.744	0.519	0.508	0.732	0.734
	Mean	0.725	0.526	0.979	0.752	0.515	0.514	0.725	0.729
MFCC	1	0.766	0.550	0.982	0.766	0.549	0.545	0.766	0.770
	2	0.761	0.541	0.982	0.761	0.560	0.545	0.761	0.762
	3	0.760	0.552	0.981	0.767	0.581	0.559	0.760	0.760
	4	0.752	0.563	0.981	0.772	0.546	0.551	0.752	0.757
	5	0.758	0.554	0.981	0.768	0.571	0.559	0.758	0.760
	Mean	0.759	0.552	0.982	0.767	0.561	0.552	0.759	0.762
MSPC	1	0.796	0.615	0.984	0.800	0.622	0.613	0.796	0.795
	2	0.791	0.580	0.983	0.781	0.637	0.602	0.791	0.790
	3	0.794	0.615	0.984	0.799	0.639	0.624	0.794	0.795
	4	0.784	0.602	0.983	0.792	0.579	0.588	0.784	0.784
	5	0.789	0.585	0.983	0.784	0.598	0.588	0.789	0.788
	Mean	0.791	0.599	0.983	0.791	0.615	0.603	0.791	0.790
LMSPC	1	0.791	0.623	0.983	0.803	0.643	0.629	0.791	0.792
	2	0.761	0.556	0.981	0.769	0.587	0.563	0.761	0.761
	3	0.758	0.562	0.981	0.771	0.564	0.560	0.758	0.760
	4	0.785	0.579	0.983	0.781	0.608	0.589	0.785	0.785
	5	0.768	0.567	0.981	0.774	0.602	0.579	0.768	0.765
	Mean	0.773	0.577	0.982	0.780	0.601	0.584	0.773	0.772
Inertial & MFCC	1	0.819	0.630	0.986	0.808	0.630	0.628	0.819	0.818
	2	0.803	0.622	0.985	0.803	0.620	0.619	0.803	0.804
	3	0.807	0.620	0.985	0.802	0.630	0.624	0.807	0.807
	4	0.809	0.621	0.985	0.803	0.640	0.628	0.809	0.808
	5	0.825	0.638	0.986	0.812	0.647	0.640	0.825	0.823
	Mean	0.813	0.626	0.985	0.806	0.633	0.628	0.813	0.812

Table 4.8: 5-Fold Results of Placement Independent Contexts (continued)

Sensor	Fold ID	ACC	REC	SPE	BACC	PRE	Mac-F1	Mic-F1	W-F1
Inertial & MSPC	1	0.791	0.597	0.984	0.790	0.609	0.599	0.791	0.792
	2	0.805	0.611	0.984	0.798	0.640	0.620	0.805	0.802
	3	0.814	0.631	0.985	0.808	0.660	0.642	0.814	0.812
	4	0.802	0.593	0.984	0.789	0.619	0.604	0.802	0.801
	5	0.814	0.660	0.986	0.823	0.659	0.657	0.814	0.816
	Mean	0.805	0.618	0.985	0.802	0.638	0.624	0.805	0.804
Inertial & LMSPC	1	0.800	0.637	0.984	0.811	0.647	0.637	0.800	0.801
	2	0.797	0.615	0.984	0.800	0.619	0.610	0.797	0.797
	3	0.806	0.615	0.985	0.800	0.649	0.625	0.806	0.807
	4	0.807	0.644	0.985	0.814	0.632	0.633	0.807	0.807
	5	0.790	0.577	0.983	0.780	0.596	0.583	0.790	0.788
	Mean	0.800	0.618	0.984	0.801	0.629	0.617	0.800	0.800

Activity-context pair based average F1 scores are visualized in figure 4.6. In general, fusion of both inertial and auditory sensors produces superior results than others, similar to the observations in other levels. Some contexts have very high overall performance while others have low overall performance due to unequal data distribution.

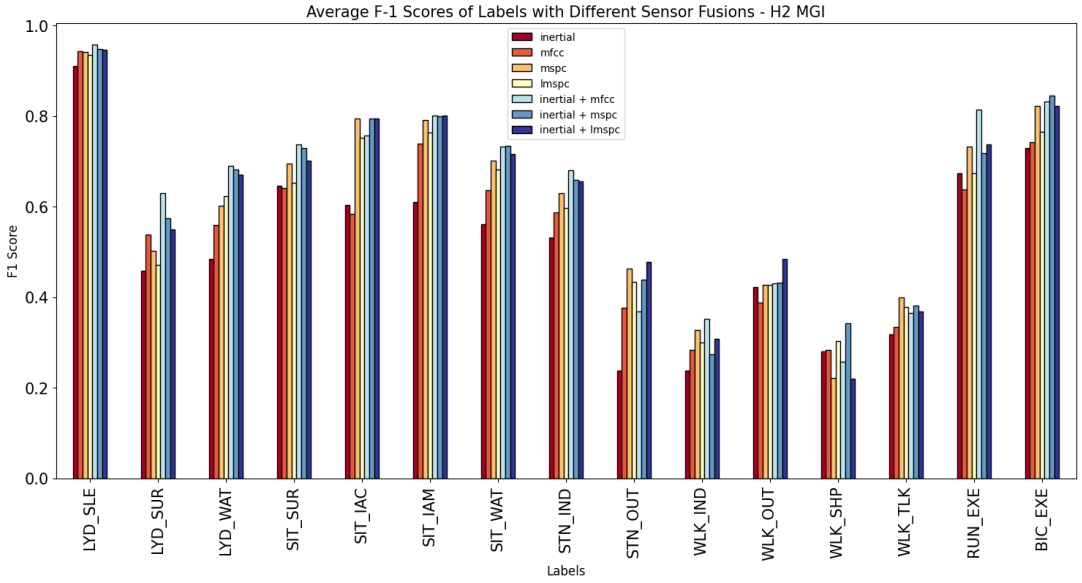


Figure 4.6: PIC Main Group Independent F1 scores per context - 5-Fold

Figure 4.6 also demonstrates importance of the audio. For example, model performs with much more higher recognition accuracy rate in case of activities including vehicle noise, such as "Sitting, In a Car" or environmental noise such as "Standing Outside".

In hierarchy level three (H3), multi-class classification is operated with 28 classes represented in the level three in Figure 4.1. In this level, smartphone placements are handled separately per each primary physical activity and context pairs. Their general name is Placement Dependent Contexts (*PDC*). Results of model performance is listed in Table 4.9 per each input type, with all fold results as well as overall result. When inertial and auditory sensor characteristics are fused, the mean findings obtained from 5-fold cross validation performance metrics show that the recognition performance improves. The model show similar performances with three versions of the fusion, with distinct audio features. Balanced accuracy score is measured as 76% in overall. Using inertial sensors only is resulted with 72% balanced accuracy score. Among three audio features, MFCC become the most informative one during prediction with score 71%.

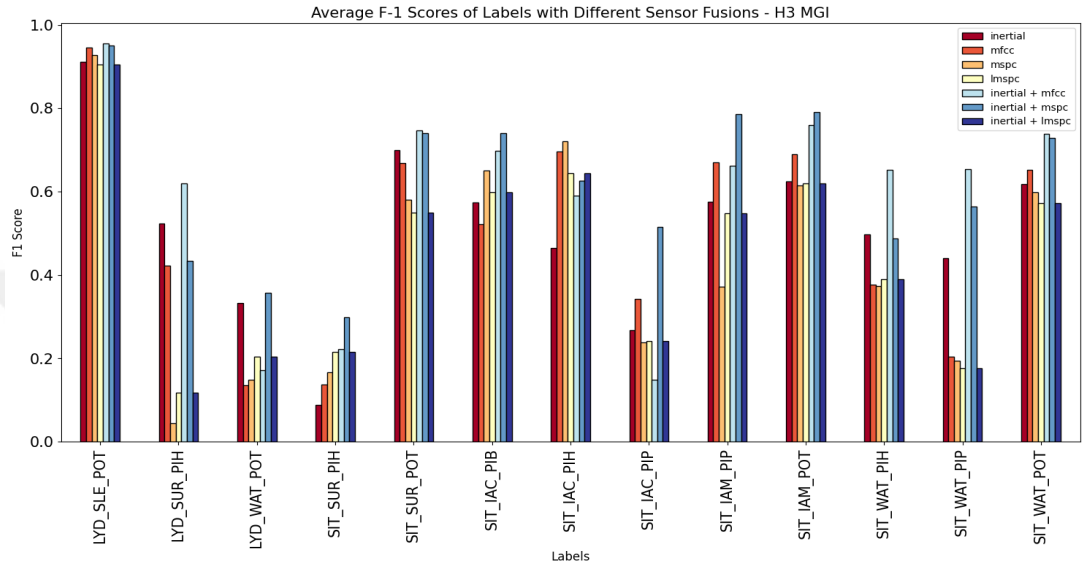
Table 4.9: 5-Fold Results of Placement Dependent Contexts

Sensor	Fold ID	ACC	REC	SPE	BACC	PRE	Mac-F1	Mic-F1	W-F1
Inertial	1	0.730	0.454	0.989	0.722	0.460	0.439	0.730	0.739
	2	0.733	0.475	0.989	0.732	0.466	0.458	0.733	0.737
	3	0.730	0.427	0.989	0.708	0.405	0.403	0.730	0.733
	4	0.712	0.404	0.988	0.696	0.385	0.383	0.712	0.716
	5	0.718	0.451	0.989	0.720	0.464	0.437	0.718	0.727
	Mean	0.725	0.442	0.989	0.716	0.436	0.424	0.725	0.731
MFCC	1	0.749	0.508	0.990	0.749	0.519	0.499	0.749	0.755
	2	0.753	0.415	0.990	0.702	0.443	0.415	0.753	0.754
	3	0.728	0.423	0.989	0.706	0.408	0.404	0.728	0.732
	4	0.759	0.412	0.990	0.701	0.466	0.430	0.759	0.758
	5	0.749	0.425	0.990	0.707	0.403	0.409	0.749	0.753
	Mean	0.748	0.436	0.990	0.713	0.448	0.431	0.748	0.750

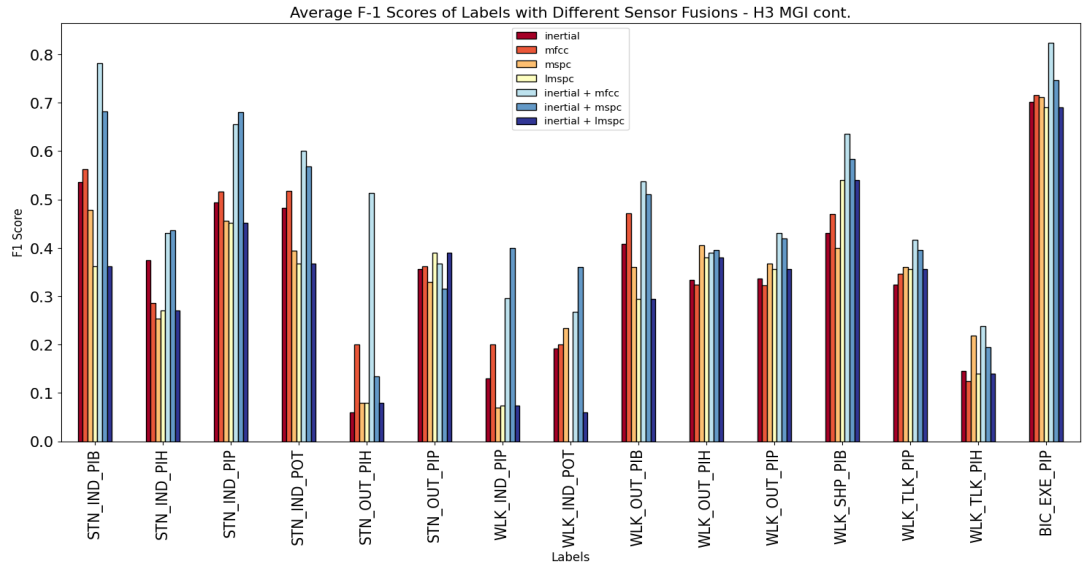
Table 4.10: 5-Fold Results of Placement Dependent Contexts (continued)

Sensor	Fold ID	ACC	REC	SPE	BACC	PRE	Mac-F1	Mic-F1	W-F1
MSPC	1	0.727	0.429	0.989	0.709	0.424	0.419	0.727	0.731
	2	0.732	0.387	0.989	0.688	0.434	0.399	0.732	0.734
	3	0.737	0.427	0.989	0.708	0.483	0.431	0.737	0.742
	4	0.717	0.394	0.988	0.691	0.405	0.394	0.717	0.720
	5	0.645	0.280	0.985	0.633	0.284	0.274	0.645	0.645
	Mean	0.712	0.384	0.988	0.686	0.406	0.384	0.712	0.714
LMSPC	1	0.681	0.410	0.987	0.698	0.430	0.411	0.681	0.687
	2	0.672	0.376	0.986	0.681	0.325	0.345	0.672	0.675
	3	0.716	0.378	0.988	0.683	0.402	0.380	0.716	0.716
	4	0.678	0.329	0.987	0.658	0.348	0.329	0.678	0.684
	5	0.689	0.453	0.987	0.720	0.413	0.416	0.689	0.702
	Mean	0.687	0.389	0.987	0.688	0.384	0.376	0.687	0.693
Inertial & MFCC	1	0.802	0.535	0.992	0.763	0.505	0.511	0.802	0.804
	2	0.798	0.535	0.992	0.764	0.529	0.524	0.798	0.803
	3	0.788	0.543	0.991	0.767	0.563	0.541	0.788	0.791
	4	0.816	0.547	0.993	0.770	0.631	0.572	0.816	0.815
	5	0.793	0.523	0.992	0.758	0.577	0.531	0.793	0.793
	Mean	0.799	0.537	0.992	0.764	0.561	0.536	0.799	0.801
Inertial & MSPC	1	0.801	0.533	0.992	0.763	0.544	0.518	0.801	0.800
	2	0.770	0.534	0.991	0.763	0.546	0.516	0.770	0.773
	3	0.810	0.570	0.992	0.781	0.607	0.575	0.810	0.812
	4	0.809	0.545	0.992	0.768	0.622	0.556	0.809	0.809
	5	0.772	0.488	0.991	0.740	0.521	0.484	0.772	0.773
	Mean	0.793	0.534	0.992	0.763	0.568	0.530	0.793	0.794
Inertial & LMSPC	1	0.787	0.558	0.991	0.775	0.572	0.556	0.787	0.791
	2	0.795	0.586	0.992	0.789	0.622	0.592	0.795	0.796
	3	0.778	0.495	0.991	0.743	0.509	0.497	0.778	0.778
	4	0.774	0.481	0.991	0.736	0.486	0.477	0.774	0.776
	5	0.774	0.540	0.991	0.765	0.597	0.545	0.774	0.778
	Mean	0.782	0.532	0.991	0.762	0.557	0.533	0.782	0.784

Two subfigures in figure 4.7 demonstrates model performance over each class for each input type, in F1 measurement unit. PDC-based performance is revealed to be very variable for each class, with certain situations doing exceptionally well. The main cause of this problem is an unequal distribution of user data. Labels with a larger sample of examples work well and model learns their features very accurately, but it performs poorly when recognizing labels with a smaller sample of data.



(a)



(b)

Figure 4.7: PDC Main Group Independent F1 scores per context-placement - 5-Fold

4.2 Hierarchical Evaluation based on Main Group Dependency

In this section, first, recognition performance of primary physical activities (H1) are evaluated. Performance of context recognition in case of placement independent contexts (H2) and placement dependent contexts (H3) are represented within groups that share same prior physical activity and at same hierarchical level. Figure 4.8 illustrates user-defined labels per each step of the hierarchy and test groups where results are listed under this section. Holdout and 5-fold cross validation evaluations are performed.

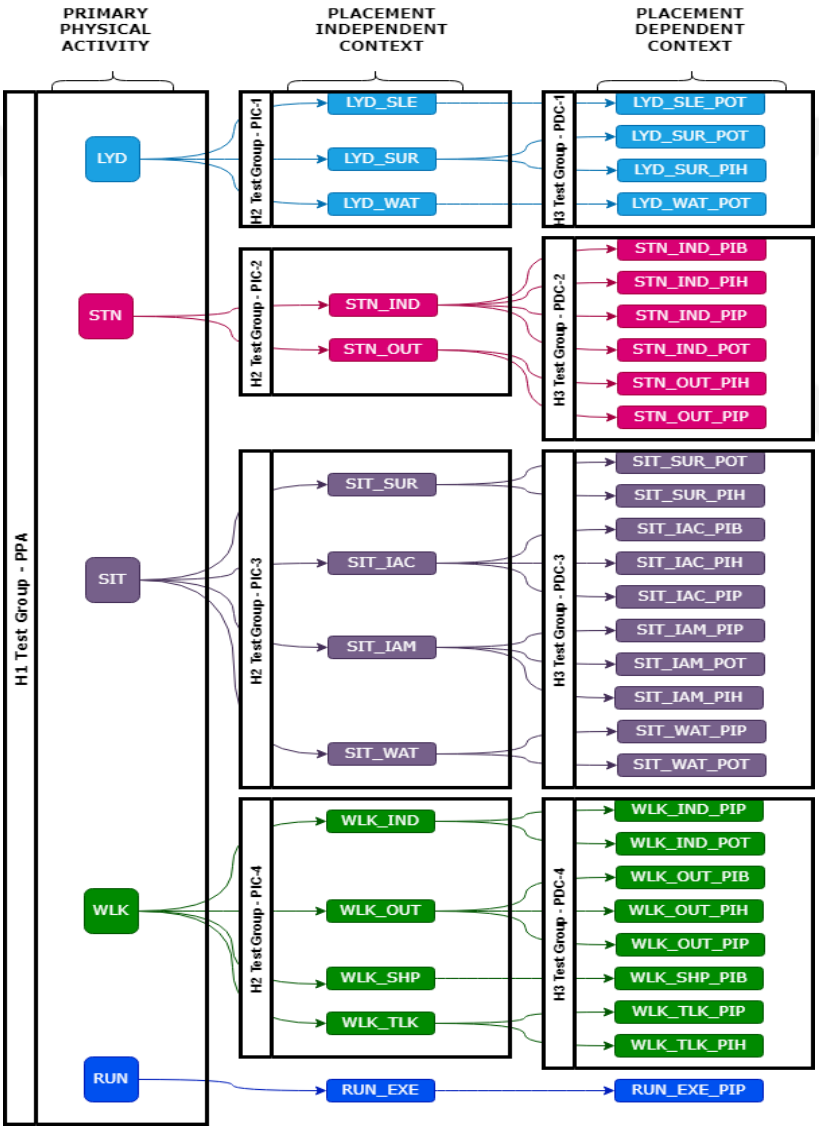


Figure 4.8: Hierarchically User-Defined Activities and Test Groups - Main Group Dependent

4.2.1 Holdout Cross Validation Results

The tests are carried out in this part using the holdout cross validation approach, which is described in greater detail above. Holdout method is applied five times and average results are listed.

Recognition performance for primary physical activity recognition task is similar to results in previous section for same hierarchy level, since instead of 6 classes, there are 5 main physical activities, "Bicycling" is not included in this classification. Figure 4.8 includes those five activities with *H1 Test Group - PPA*. Scores of each iteration, as well as mean for different metrics are listed in Table 4.11. Overall BA score measured with only inertial sensor features is 82%. However, it is observed from the results that only using audio features whether it is MFCC or MSPC, recognition performance become nearly 83%, slightly more than inertial only. This concludes audio only could be used to predict primary human activity. As predicted, audio and inertial sensor fusion images have an empowering effect with the described lightweight model. Average of balanced accuracy scores for MSPC and four inertial sensor fusion is obtained as 90.7% where MFCC and inertial combination gives BA rate 90%.

Figure 4.9 demonstrates model performance over five primary physical contexts, with F1 score unit. The overall performance is much improved by combining three characteristics of each inertial sensor and several feature configurations of the audio sensor. The model especially confuses "Standing" and "Walking" activities with each other. Furthermore, in the case of "Running" physical activity, the auditory signal becomes significantly more informative.

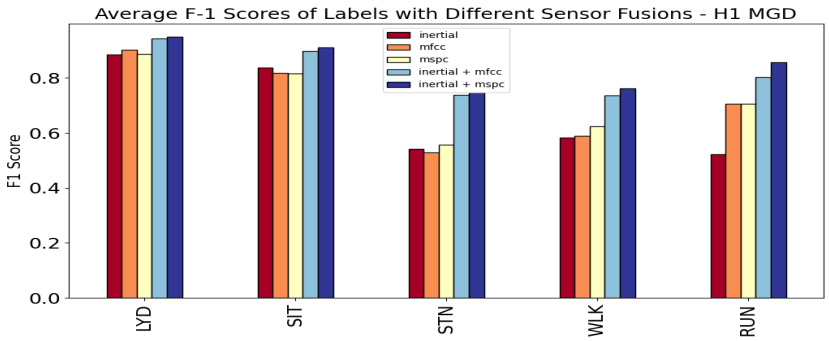


Figure 4.9: PPA Main Group Dependent F1 scores per activity - Holdout

Table 4.11: Holdout Results of Primary Physical Activity Results

Sensor	Holdout ID	ACC	REC	SPE	BACC	PRE	Mac-F1	Mic-F1	W-F1
Inertial	1	0.795	0.702	0.940	0.821	0.656	0.672	0.795	0.797
	2	0.788	0.692	0.938	0.815	0.647	0.663	0.788	0.790
	3	0.800	0.718	0.941	0.830	0.664	0.683	0.800	0.802
	4	0.799	0.693	0.941	0.817	0.655	0.669	0.799	0.801
	5	0.806	0.711	0.943	0.827	0.665	0.682	0.806	0.807
	Mean	0.798	0.703	0.941	0.822	0.657	0.674	0.798	0.800
MFCC	1	0.791	0.708	0.937	0.823	0.704	0.705	0.791	0.792
	2	0.797	0.714	0.939	0.827	0.710	0.712	0.797	0.798
	3	0.802	0.726	0.941	0.834	0.717	0.721	0.802	0.803
	4	0.796	0.705	0.939	0.822	0.706	0.705	0.796	0.798
	5	0.800	0.720	0.940	0.830	0.742	0.729	0.800	0.802
	Mean	0.797	0.715	0.939	0.827	0.716	0.714	0.797	0.799
MSPC	1	0.793	0.727	0.938	0.833	0.714	0.720	0.793	0.793
	2	0.795	0.717	0.939	0.828	0.713	0.715	0.795	0.798
	3	0.803	0.727	0.942	0.834	0.720	0.723	0.803	0.804
	4	0.801	0.721	0.940	0.831	0.718	0.719	0.801	0.802
	5	0.786	0.723	0.936	0.830	0.709	0.715	0.786	0.787
	Mean	0.796	0.723	0.939	0.829	0.715	0.719	0.796	0.797
Inertial & MFCC	1	0.892	0.844	0.969	0.907	0.824	0.833	0.892	0.892
	2	0.886	0.828	0.966	0.897	0.818	0.823	0.886	0.886
	3	0.879	0.818	0.964	0.891	0.806	0.812	0.879	0.879
	4	0.890	0.834	0.968	0.901	0.821	0.827	0.890	0.890
	5	0.893	0.836	0.969	0.903	0.823	0.829	0.893	0.894
	Mean	0.889	0.832	0.967	0.900	0.818	0.825	0.889	0.889
Inertial & MSPC	1	0.894	0.847	0.969	0.908	0.857	0.852	0.894	0.896
	2	0.885	0.826	0.967	0.896	0.822	0.824	0.885	0.887
	3	0.900	0.853	0.971	0.912	0.844	0.848	0.900	0.902
	4	0.898	0.850	0.970	0.910	0.850	0.850	0.898	0.900
	5	0.899	0.843	0.970	0.906	0.873	0.856	0.899	0.901
	Mean	0.895	0.844	0.969	0.907	0.849	0.846	0.895	0.897

User-defined activities in the second and third steps of the hierarchy tree, placement independent contexts and placement dependent contexts respectively, are used to per-

form model evaluation with user-defined activities in the branches of the primary activity they are connected and that share same hierarchical layers. Therefore, for each level, there are four different classification are conducted.

Table 4.12: Average Holdout Results of Placement Independent Context Group Testing

Test Group	Sensor	ACC	REC	SPE	BACC	PRE	Mac-F1	Mic-F1	W-F1
PIC-1	Inertial	0.965	0.765	0.902	0.834	0.756	0.759	0.965	0.966
	MFCC	0.970	0.784	0.912	0.848	0.791	0.786	0.970	0.970
	MSPC	0.965	0.753	0.895	0.824	0.761	0.754	0.965	0.965
	Inertial & MFCC	0.974	0.842	0.940	0.891	0.804	0.821	0.974	0.973
	Inertial & MSPC	0.973	0.798	0.920	0.859	0.795	0.794	0.973	0.973
PIC-2	Inertial	0.938	0.770	0.799	0.790	0.729	0.755	0.938	0.937
	MFCC	0.948	0.679	0.679	0.679	0.720	0.694	0.948	0.946
	MSPC	0.949	0.694	0.694	0.694	0.735	0.707	0.949	0.948
	Inertial & MFCC	0.952	0.825	0.825	0.825	0.752	0.791	0.952	0.952
	Inertial & MSPC	0.950	0.799	0.799	0.799	0.740	0.766	0.950	0.951
PIC-3	Inertial	0.786	0.760	0.920	0.840	0.777	0.767	0.786	0.786
	MFCC	0.777	0.768	0.919	0.844	0.761	0.762	0.777	0.777
	MSPC	0.809	0.797	0.927	0.862	0.813	0.803	0.809	0.808
	Inertial & MFCC	0.974	0.842	0.940	0.891	0.804	0.821	0.974	0.974
	Inertial & MSPC	0.838	0.836	0.938	0.887	0.848	0.840	0.838	0.838
PIC-4	Inertial	0.493	0.469	0.810	0.637	0.470	0.463	0.493	0.492
	MFCC	0.524	0.510	0.826	0.668	0.499	0.493	0.524	0.521
	MSPC	0.464	0.389	0.797	0.593	0.404	0.389	0.464	0.460
	Inertial & MFCC	0.519	0.481	0.820	0.651	0.480	0.475	0.519	0.517
	Inertial & MSPC	0.569	0.542	0.843	0.692	0.539	0.535	0.569	0.566

For each test group, model is tested with different test data five times. Average values of those with respect to corresponding test group is listed with Table 4.12 for position independent human activity recognition. At this level, each activity indicates a combination of a context and a primary physical activity. Test group *PIC-1* shows how proposed model perform predicting contexts for primary physical activity "Lying Down". Because of the dominance of context "Sleeping", this comparison is done by very imbalanced data. However, by avoiding this perplexing predicament, the model has produced efficient recognition results by giving a class weight to each class dur-

ing training. Audio and inertial sensor fusion produces more accurate data than any sensor on its own. The combination of inertial and MFCC yields an 89% balanced accuracy score, whereas the combination of inertial and MSPC yields an 85%. In this situation, audio feature MFCC, has significantly improved recognition performance compared to MSPC. There is another observation in this case is that, accuracy results are very high compared to balanced accuracy. Accuracy gives true predictions over all, therefore, it is possible to conclude that model is successful in estimating the class with a considerably higher number of examples.

Test group *PIC-2* represents, two contexts of "Standing". Contributing sample image count of "Indoors" is excessively much compared to "Outside". Again, audio existence in the system rises performance score. Inertial only performs with 79% balanced accuracy rate, while inertial and MSPC fusion follows it with nearly 80%. In this case, MFCC is the key audio feature which increases overall performance combined with inertial sensors. Average balanced accuracy result become 82.5%. While neither type of audio helps to predict as accurately as inertial alone, when used as a fusion with inertials, they have an improving effect on system performance.

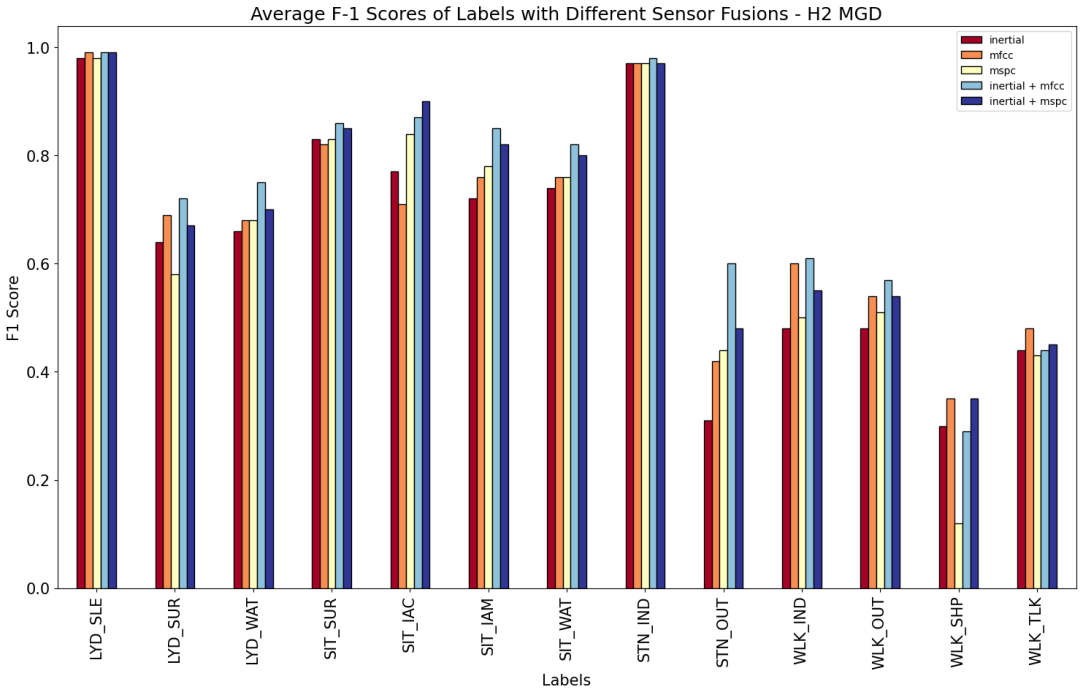
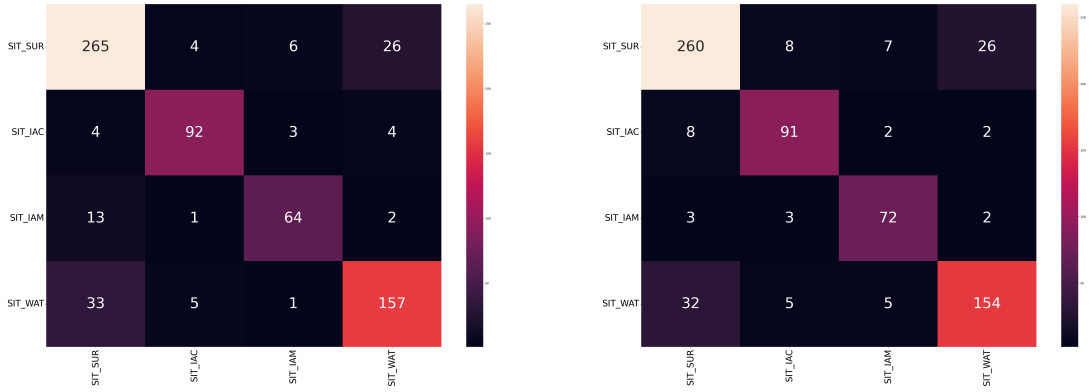


Figure 4.10: PIC Main Group Dependent F1 scores per context - Holdout

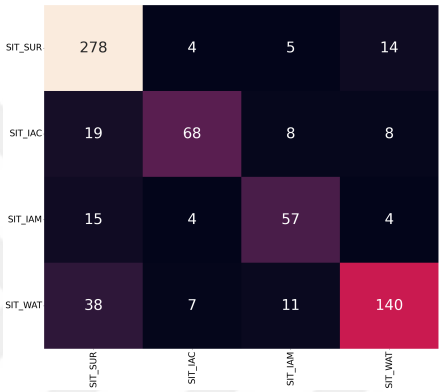
Model performance to distinct different contexts based on provided data is shown with the figure 4.10. Values of each context here illustrates context-based performance within corresponding test group, in *PIC* groups. Average F1 scores of five iterations are shown. In general, sensor fusion approach works well compared to single sensory information. Context "Sleeping" have pretty high recognition performance compared to others, but this cluster includes extremely high sample size compared to others. Still, model could utilize images to yield good performance over small sample sized contexts as well, by using class weight technique. *PIC-1*, *PIC2* and *PIC-3* yields great performances, average balanced accuracy 87% with sensor fusion, with MFCC features. However, model could not recognize *PIC-4*, activity "Walking" contexts as well as other groups in the hierarchy, which results BA score of 84% overall.

Placement Independent Context group *PIC-3* indicates contexts from ancestor physical activity "Sitting". There are four contexts to be predicted in this case. In case of sample sizes, there is a more accurate distribution between class samples, therefore, class weights. Average balanced accuracy with only using inertial sensors is around 84%. Sensor fusion with both MFCC and MSPC performs with 89% and 88.6% respectively. However, there is an interesting outcome observed in this case. Only MSPC's success in correctly identifying which of these four classes any image supplied to it belongs to is 86%, little more than only inertial. Contexts such as "In a car", "In a meeting" belongs to that class is more predictable with given either environmental sounds or human interaction. Figure 4.11 illustrates confusion matrices per each sensor modality.

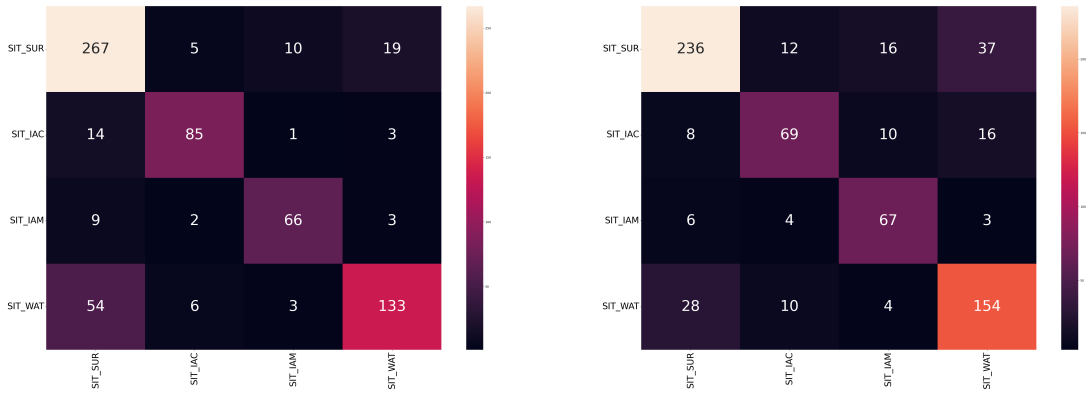
Last test group is in the second hierarchy level is *PIC-4*, contexts of activity "Walking". Among all other test groups in this hierarchy level, lower results are obtained in this group. When possible issues are investigated, it is observed that system could properly distinguish between contexts "Outside" and "Talking". A reason might be that environmental sounds in an external setting where people are communicating with one other could also generate difficulty in the system's capacity to recognize context, even if the person is not speaking. All in all, best performance is achieved with sensor fusion again, with MSPC and inertial sensors as BA 69%.



(a) (b)



(c)



(d) (e)

Figure 4.11: H2 Test Group - Placement Independent Context-3 Confusion Matrices, iter 4 (a) Inertial & MSPC fusion (b) Inertial & MFCC fusion (c) Inertial (d) MSPC (e) MFCC - Holdout

In hierarchy level three, activity groupings divided into different contexts in the previous level are also grouped according to smartphone placements. Table 4.13 denotes the results of four different category of user-defined activities.

At level three hierarchy, a primary physical activity occur together with a context which is represented with multiple input types, based on smartphone placement. First test group is called *PDC-1*, which indicates primary activity "Lying Down"'s context classes constructed by different settlement of smartphone. In this test group, sensor fusion had a enhancing impact. Model performs with 83% balanced accuracy rate while using inertial sensory information during prediction. When audio sensors are included, the success rate rises to 84-85% as BA. As could be predicted from accuracy rates, it is a highly imbalanced sub-group.

Table 4.13: Average Holdout Results of Placement Dependent Context Group Testing

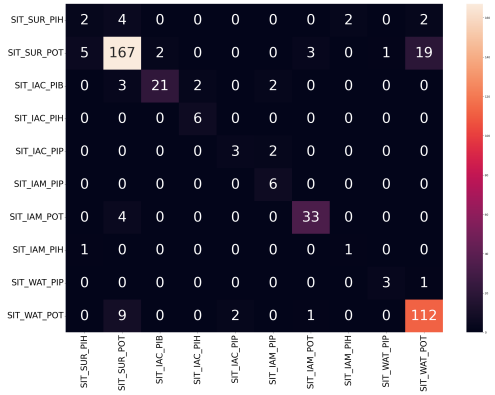
Test Group	Sensor	ACC	REC	SPE	BACC	PRE	Mac-F1	Mic-F1	W-F1
PDC-1	Inertial	0.979	0.730	0.934	0.832	0.769	0.736	0.979	0.980
	MFCC	0.977	0.686	0.926	0.806	0.743	0.695	0.977	0.977
	MSPC	0.971	0.572	0.900	0.736	0.630	0.592	0.971	0.972
	Inertial & MFCC	0.973	0.747	0.934	0.840	0.699	0.721	0.973	0.974
	Inertial & MSPC	0.972	0.794	0.926	0.860	0.878	0.822	0.972	0.973
PDC-2	Inertial	0.771	0.638	0.943	0.801	0.686	0.648	0.771	0.771
	MFCC	0.724	0.534	0.929	0.731	0.617	0.551	0.724	0.726
	MSPC	0.642	0.470	0.923	0.701	0.467	0.468	0.642	0.642
	Inertial & MFCC	0.870	0.835	0.967	0.901	0.873	0.839	0.870	0.871
	Inertial & MSPC	0.830	0.673	0.956	0.815	0.828	0.713	0.830	0.830
PDC-3	Inertial	0.742	0.574	0.966	0.770	0.513	0.536	0.742	0.743
	MFCC	0.780	0.581	0.968	0.774	0.605	0.581	0.780	0.781
	MSPC	0.784	0.595	0.969	0.782	0.625	0.603	0.784	0.784
	Inertial & MFCC	0.834	0.680	0.976	0.828	0.698	0.678	0.834	0.835
	Inertial & MSPC	0.845	0.744	0.978	0.861	0.682	0.703	0.845	0.845
PDC-4	Inertial	0.506	0.562	0.924	0.743	0.531	0.527	0.506	0.507
	MFCC	0.466	0.505	0.917	0.711	0.520	0.491	0.466	0.466
	MSPC	0.322	0.269	0.893	0.581	0.323	0.261	0.322	0.323
	Inertial & MFCC	0.573	0.612	0.932	0.772	0.614	0.607	0.573	0.572
	Inertial & MSPC	0.476	0.509	0.917	0.713	0.561	0.494	0.476	0.477

In second test group, *PDC-2*, group of *Standing* physical activity's placement dependent contexts are evaluated. With utilizing inertial sensor color coded images, model performance reaches 80% BA in average. Audio MSPC and inertial combination rises that performance to 81%, where MFCC feature usage increase the performance to nearly 90%. Although not a particularly impressive result in single sensor, the use of sensors in fusion improves the overall performance.

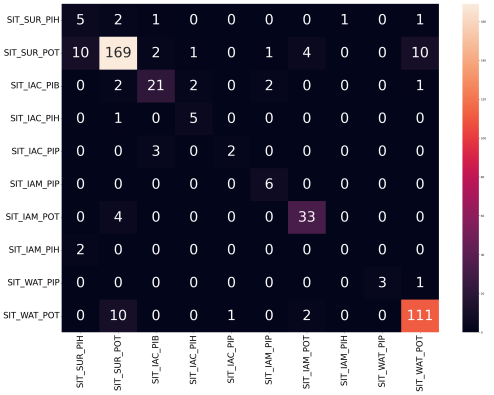
PDC-3 indicates group of contexts derived with respect to main activity "Sitting". The effect of numerous sensors combined together is seen here, as it is in the others. Model trained with only inertial sensors have 77% accurate prediction rate in case of BA. With the help of audio sensors, performance score rises to 86% in average using MSPC where it is observed around 82% for MFCC. These performance-enhancing results highlight the importance of the audio sensor. It could be seen from confusion matrices of the sensors. In figure 4.12, it is seen that from inertial sensor only confusion matrix, having the same primary physical activity "Sitting" with the same smartphone placement "Phone on Table", model is vunerable to predict whether image belongs to context class "Surfing the Internet" or "Watching TV". However, with the help of using audio sensors, this vunerability decays during prediction in the same situation.

Lastly, placement oriented user-defined activities derived from primary activity "Walking" which represented with group *PDC-4* are evaluated. Deep neural network performance trained by inertial sensor individually gives 74% balanced accuracy score. The overall balanced accuracy score decreases when inertial sensors combined with MSPC. On the other hand, performance of the model tend to increase by using MFCC features of audio rather than MSPC. Performance rises to 77% by using MFCC features and inertial sensors together. Deep model performance trained only by those audio features become 71%, i.e. system accurately tell primary physical activity, context and smartphone placement of a person with 71% by only looking at sounds around.

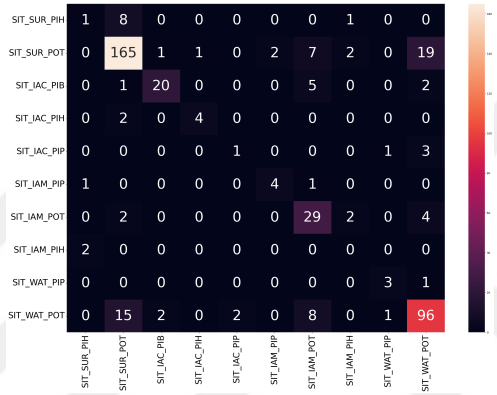
Figure 4.13 illustrates F1 score yielded by model based on context-placement pairs. MFCC, outperforms MSPC at this level for some user-defined labels. For example,



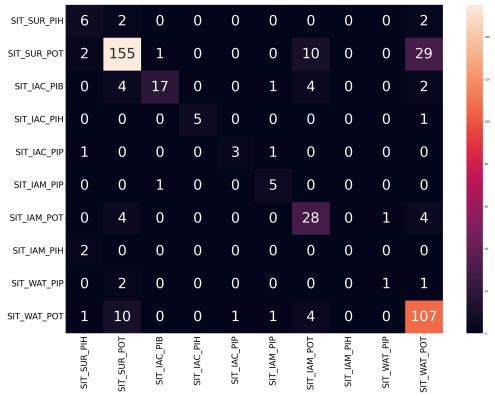
(a)



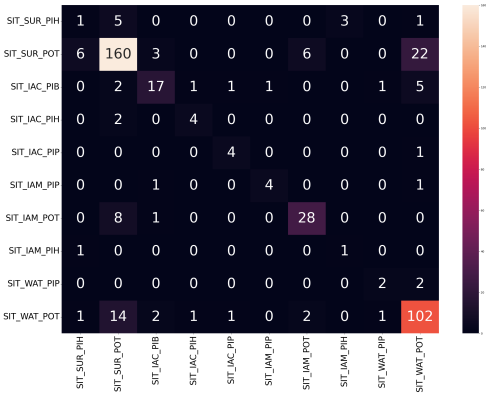
(b)



(c)



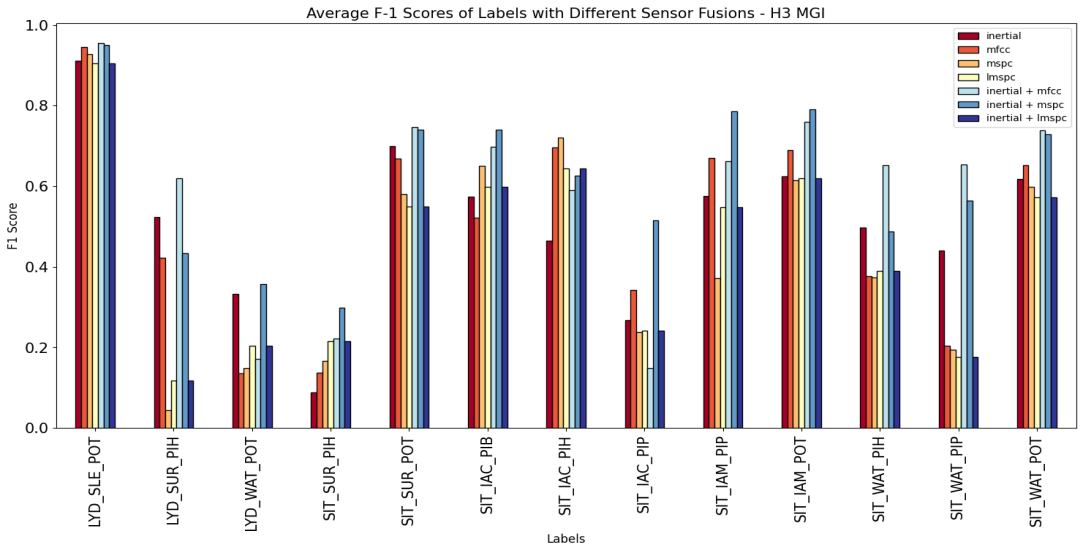
(d)



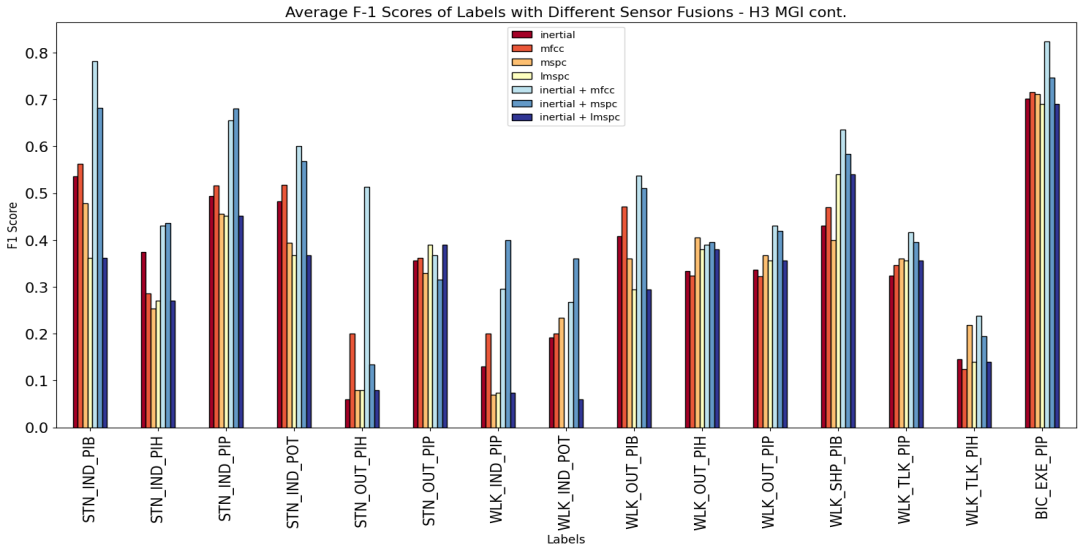
(e)

Figure 4.12: H3 Test Group - Placement Independent Context-3 Confusion Matrices, iter 2 (a) Inertial & MSPC fusion (b) Inertial & MFCC fusion (c) Inertial (d) MSPC (e) MFCC - Holdout

MFCCs are particularly potent compared to MSPCs in the placement-dependent context "Sitting, In a Meeting, Phone in Hand," and as a result, MFCC and inertial fusion performance. In case of BA score, average performance of test groups *PDC-1*, *PDC-2* and *PDC-3* 87%, as observed for the same main group dependent part at the previous level. However, in hierarchy three, overall performance of *PDC-4* is increased to 77%. The more successful outcome emphasizes the importance of phone placement.



(a)



(b)

Figure 4.13: PDC Main Group Dependent F1 scores per context-placement - Holdout

4.2.2 5-Fold Cross Validation Results

The tests are carried out in this section using a 5-fold cross validation procedure, as detailed above. The findings are generated using each part as test data and the mean results of them are reported.

Recognizing operations are done for five main physical activities with all inertial, audio, and their combination data at hierarchy level one (H1). Figure 4.8 shows which classes are used in the classification. Table 4.14 lists results with different metrics. Model performs with balanced accuracy score 82%. Among three audio features tried here, model gives the most accurate results with MSPC feature of audio. Other audio features, MFCC and LMSPC, also contribute by themselves with values that are extremely near to the inertial sensors' score. However, the model shows great performance, 90% with audio and inertial sensor fusion. The highest average balanced accuracy score is obtained by using inertial and MSPC sensor fusion, where MFCC and inertial fusion result is very score to that. Although the score of LMSPC and inertial combination is lower than the others, as a result, it is also observed that better results are achieved if sensors are used together.

Even though "Sitting" and "Lying down" activities have large data contribution, still very small data seem to be efficient while detecting activities as could be seen from "Running" in figure 4.14.

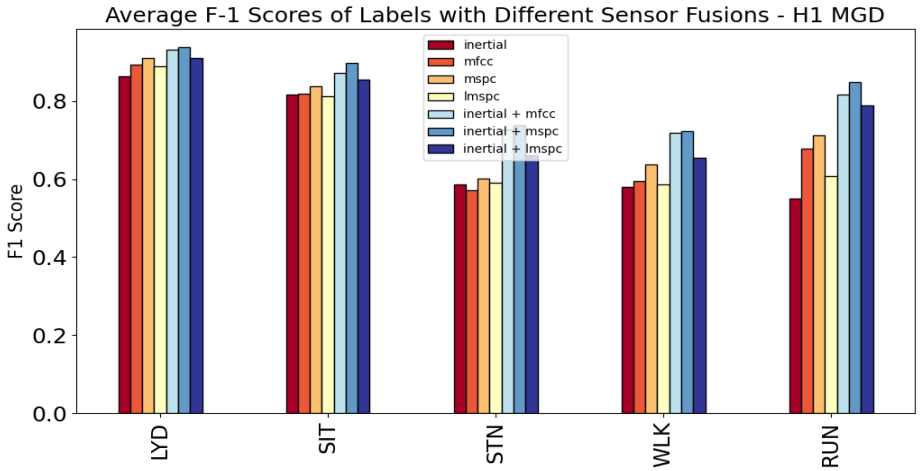


Figure 4.14: PPA Main Group Dependent F1 scores per activity - 5-Fold

Table 4.14: 5-Fold Results of Primary Physical Activities

Sensor	Fold ID	ACC	REC	SPE	BACC	PRE	Mac-F1	Mic-F1	W-F1
Inertial	1	0.789	0.713	0.938	0.825	0.659	0.681	0.789	0.791
	2	0.779	0.692	0.934	0.813	0.656	0.672	0.779	0.781
	3	0.789	0.699	0.938	0.819	0.654	0.673	0.789	0.791
	4	0.787	0.712	0.937	0.825	0.673	0.690	0.787	0.789
	5	0.783	0.706	0.936	0.821	0.668	0.685	0.783	0.784
	Mean	0.785	0.705	0.937	0.821	0.662	0.680	0.785	0.787
MFCC	1	0.789	0.691	0.937	0.814	0.716	0.701	0.789	0.790
	2	0.799	0.701	0.940	0.821	0.727	0.712	0.799	0.800
	3	0.800	0.703	0.941	0.822	0.727	0.713	0.800	0.801
	4	0.795	0.697	0.940	0.818	0.732	0.711	0.795	0.797
	5	0.799	0.706	0.941	0.824	0.738	0.718	0.799	0.801
	Mean	0.796	0.700	0.940	0.820	0.728	0.711	0.796	0.798
MSPC	1	0.803	0.736	0.940	0.838	0.730	0.732	0.803	0.803
	2	0.809	0.748	0.944	0.846	0.718	0.732	0.809	0.812
	3	0.819	0.729	0.947	0.838	0.734	0.731	0.819	0.820
	4	0.827	0.747	0.949	0.848	0.773	0.757	0.827	0.828
	5	0.829	0.730	0.949	0.840	0.773	0.746	0.829	0.830
	Mean	0.817	0.738	0.946	0.842	0.746	0.740	0.817	0.818
LMSPC	1	0.790	0.675	0.938	0.806	0.707	0.685	0.790	0.792
	2	0.791	0.689	0.939	0.814	0.713	0.698	0.791	0.793
	3	0.803	0.705	0.942	0.824	0.726	0.712	0.803	0.805
	4	0.805	0.706	0.943	0.825	0.728	0.713	0.805	0.807
	5	0.777	0.652	0.933	0.792	0.700	0.673	0.777	0.776
	Mean	0.793	0.686	0.939	0.812	0.715	0.696	0.793	0.795
Inertial & MFCC	1	0.871	0.822	0.962	0.892	0.834	0.828	0.871	0.871
	2	0.882	0.837	0.966	0.901	0.845	0.841	0.882	0.882
	3	0.882	0.810	0.966	0.888	0.842	0.824	0.882	0.882
	4	0.892	0.853	0.969	0.911	0.845	0.849	0.892	0.892
	5	0.868	0.798	0.962	0.880	0.802	0.800	0.868	0.869
	Mean	0.879	0.824	0.965	0.894	0.834	0.828	0.879	0.879

Table 4.15: 5-Fold Results of Primary Physical Activities (continued)

Sensor	Fold ID	ACC	REC	SPE	BACC	PRE	Mac-F1	Mic-F1	W-F1
Inertial & MSPC	1	0.859	0.834	0.959	0.897	0.812	0.822	0.859	0.861
	2	0.867	0.862	0.962	0.912	0.824	0.841	0.867	0.869
	3	0.850	0.805	0.957	0.881	0.734	0.763	0.850	0.852
	4	0.862	0.829	0.961	0.895	0.801	0.814	0.862	0.864
	5	0.869	0.839	0.963	0.901	0.810	0.823	0.869	0.871
	Mean	0.861	0.834	0.960	0.897	0.796	0.813	0.861	0.863
Inertial & LMSPC	1	0.823	0.764	0.949	0.857	0.752	0.755	0.823	0.824
	2	0.827	0.799	0.951	0.875	0.765	0.780	0.827	0.830
	3	0.836	0.789	0.953	0.871	0.767	0.776	0.836	0.838
	4	0.839	0.809	0.954	0.882	0.776	0.791	0.839	0.841
	5	0.838	0.769	0.954	0.861	0.767	0.766	0.838	0.841
	Mean	0.833	0.786	0.952	0.869	0.765	0.774	0.833	0.835

Model evaluation is performed with user-defined activities in the branches of the primary activity they are connected to and that share the same hierarchical layers with user-defined activities in the second and third steps of the hierarchy tree. As a result, four distinct classifications are undertaken for each level.

In hierarchy level two (H2), in case of main group dependent testing, there are four main test groups, which could be seen in figure 4.8. Table 4.16 lists the results. Test group *PIC-1*, indicates contexts of "Lying down". Even though has a very unbalanced dataset, still promising results are obtained. For MFCC and inertial fusion, the balanced accuracy score of the model become 90%. In this case, MFCC only shows 3% greater performance compared to inertial sensors. It proves effect of the audio.

For test group *PIC-2*, contexts of "Standing" activity, the most approximate recognition results are achieved with inertial and MSPC fusion with score 81%. This outcome emphasizes the importance of environmental sounds in recognizing human activity, since two contexts are used here, "Indoors" and "Outside". Despite the fact that the number of samples in the context "Indoors" is significantly more than in the context "Outside," the model still learns from this unbalanced dataset using the class weight strategy.

Table 4.16: Average 5-Fold Results of Placement Independent Context Group Testing

Test Group	Sensor	ACC	REC	SPE	BACC	PRE	Mac-F1	Mic-F1	W-F1
PIC-1	Inertial	0.966	0.778	0.910	0.844	0.758	0.766	0.966	0.967
	MFCC	0.974	0.815	0.927	0.871	0.812	0.813	0.974	0.974
	MSPC	0.969	0.767	0.903	0.835	0.784	0.774	0.969	0.969
	LMSPC	0.966	0.729	0.881	0.805	0.777	0.749	0.966	0.966
	Inertial & MFCC	0.980	0.864	0.948	0.906	0.850	0.856	0.980	0.980
	Inertial & MSPC	0.976	0.824	0.928	0.876	0.829	0.824	0.976	0.976
	Inertial & LMSPC	0.975	0.808	0.925	0.867	0.816	0.810	0.975	0.975
PIC-2	Inertial	0.956	0.729	0.729	0.729	0.763	0.742	0.956	0.955
	MFCC	0.965	0.734	0.734	0.734	0.835	0.769	0.965	0.962
	MSPC	0.953	0.722	0.722	0.722	0.737	0.728	0.953	0.952
	LMSPC	0.954	0.686	0.686	0.686	0.748	0.707	0.954	0.950
	Inertial & MFCC	0.963	0.801	0.801	0.801	0.802	0.799	0.963	0.963
	Inertial & MSPC	0.968	0.814	0.814	0.814	0.833	0.820	0.968	0.967
	Inertial & LMSPC	0.968	0.761	0.761	0.761	0.848	0.797	0.968	0.965
PIC-3	Inertial	0.798	0.779	0.924	0.852	0.793	0.785	0.798	0.798
	MFCC	0.807	0.793	0.928	0.861	0.803	0.797	0.807	0.807
	MSPC	0.833	0.822	0.935	0.879	0.845	0.831	0.833	0.832
	LMSPC	0.823	0.818	0.933	0.876	0.821	0.818	0.823	0.822
	Inertial & MFCC	0.870	0.869	0.951	0.910	0.869	0.869	0.870	0.870
	Inertial & MSPC	0.866	0.863	0.949	0.906	0.875	0.868	0.866	0.866
	Inertial & LMSPC	0.865	0.867	0.949	0.908	0.869	0.867	0.865	0.865
PIC-4	Inertial	0.515	0.482	0.819	0.650	0.489	0.474	0.515	0.514
	MFCC	0.535	0.509	0.826	0.668	0.519	0.503	0.535	0.529
	MSPC	0.509	0.441	0.814	0.627	0.464	0.438	0.509	0.499
	LMSPC	0.499	0.421	0.809	0.615	0.468	0.428	0.499	0.491
	Inertial & MFCC	0.559	0.536	0.836	0.686	0.519	0.517	0.559	0.551
	Inertial & MSPC	0.564	0.539	0.831	0.685	0.518	0.524	0.564	0.558
	Inertial & LMSPC	0.559	0.537	0.837	0.687	0.524	0.522	0.559	0.554

PIC-3 indicates contexts of "Sitting" physical activity. All three version of the audio data demonstrates good results when combining with inertial sensor data, all of them have 91% balanced accuracy score. The model shows 85% by using only inertial data. The model performance is improved compared to using only MSPC audio feature, since score is measured as 88%. Confusion matrices of *PIC-3* are in figure 4.16.

Lastly, *PIC-4* group is tested. It denotes contexts of "Walking". The most confusing test group for the model in this hierarchy is this one. The overall performance is best with LMSPC and inertial sensor combination, with 69%. As observed for the holdout case, the model does not distinguish between "Outdoors" and "Talking" activities as well as others. The cause is most likely due to the person mistaking external human speech sounds for his or her own voice.

Context-based F1 score achievements are shown in figure 4.15. The model demonstrates great performance with test groups *PIC-1*, *PIC-2* and *PIC-3*. Their average balanced accuracy score is 87%. Meanwhile, in case of *PIC-4*, even though audio has a very improving effect, still scores are not high as other test groups. Still, in overall, grouping technique in hierarchical levels shows slightly better performance compared to not grouping, based on the results.

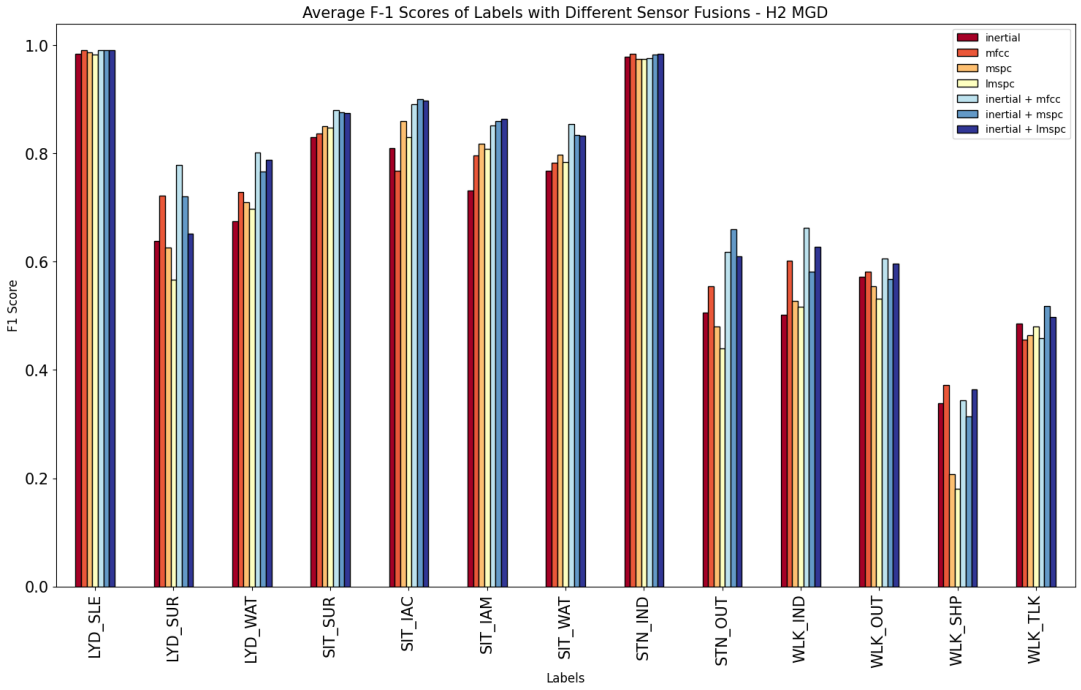


Figure 4.15: PIC Main Group Dependent F1 scores per context - 5-Fold

In hierarchy level three (H3), for main group dependent test mechanism, again, four test group is created. However, the user-defined labels are distinct from each other in another way at this level, again activity-context division but additional on smartphone placement. Table 4.17 lists average results of 5-fold cross-validation.

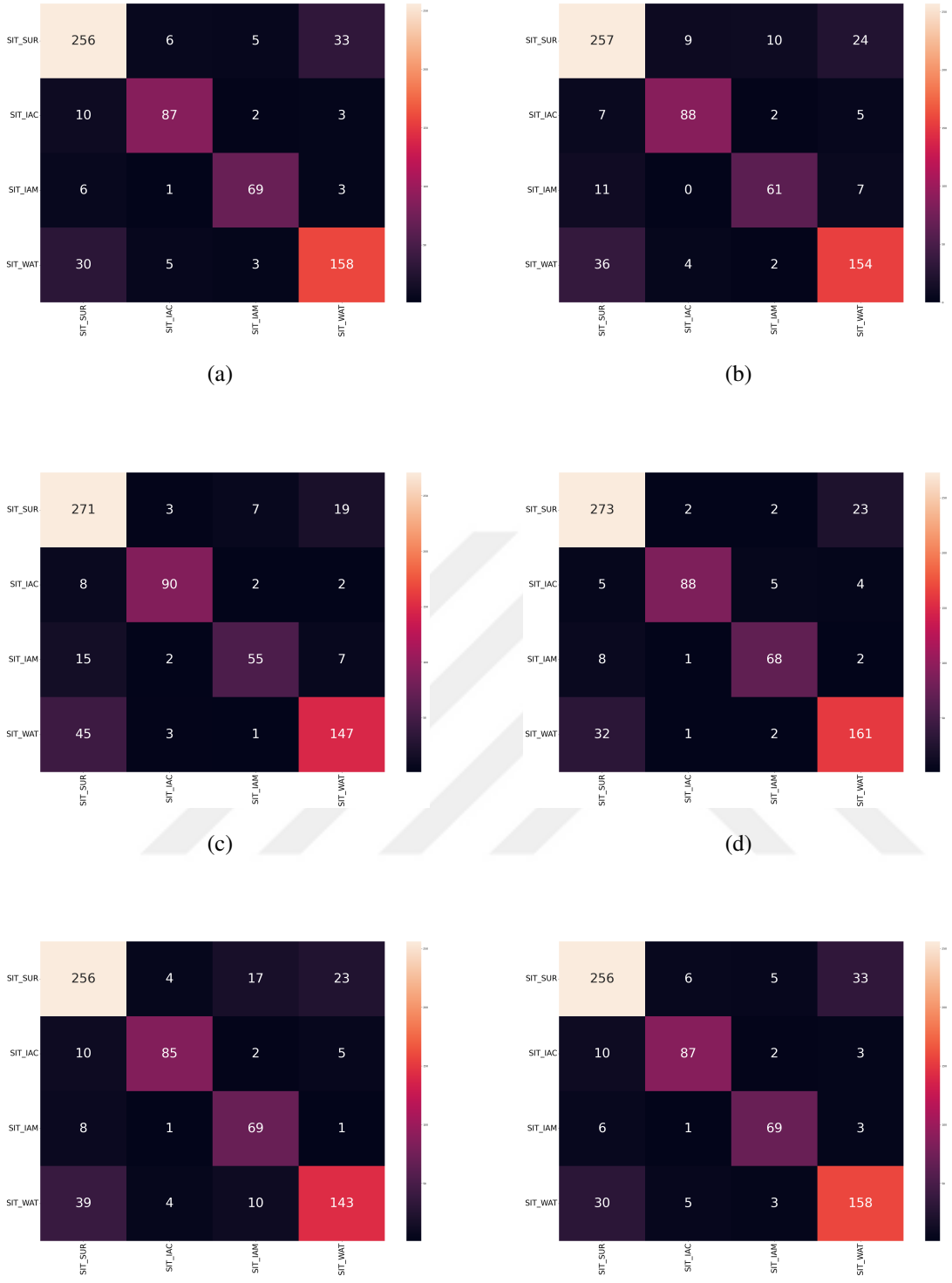


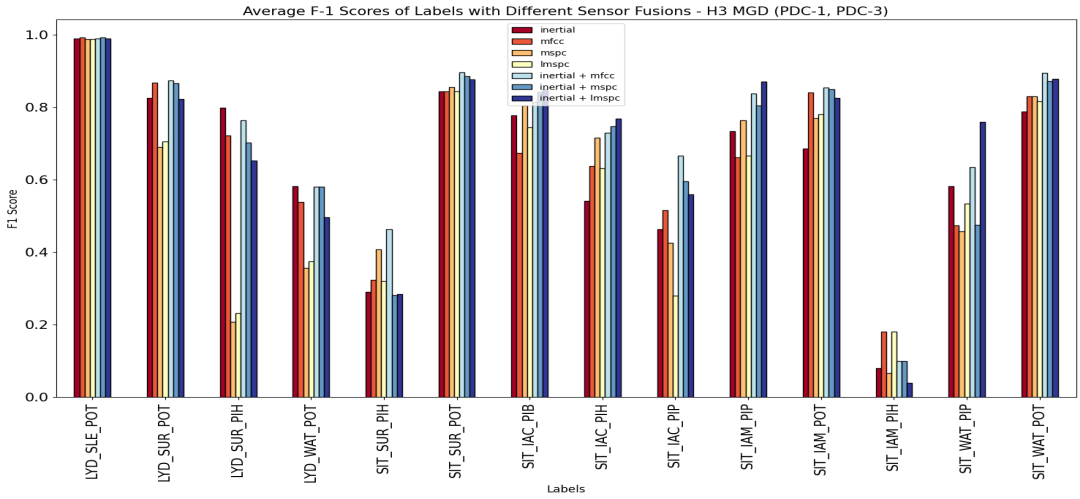
Figure 4.16: H2 Test Group - Placement Dependent Context-3 Confusion Matrices, Fold 1 (a-b) MFCC, Inertial & MFCC (c-d) MSPC, Inertial&MSPC, (e-f) LMSPC, Inertial&LMSPC, - 5-Fold

Table 4.17: Average 5-Fold Results of Placement Dependent Context Group Testing

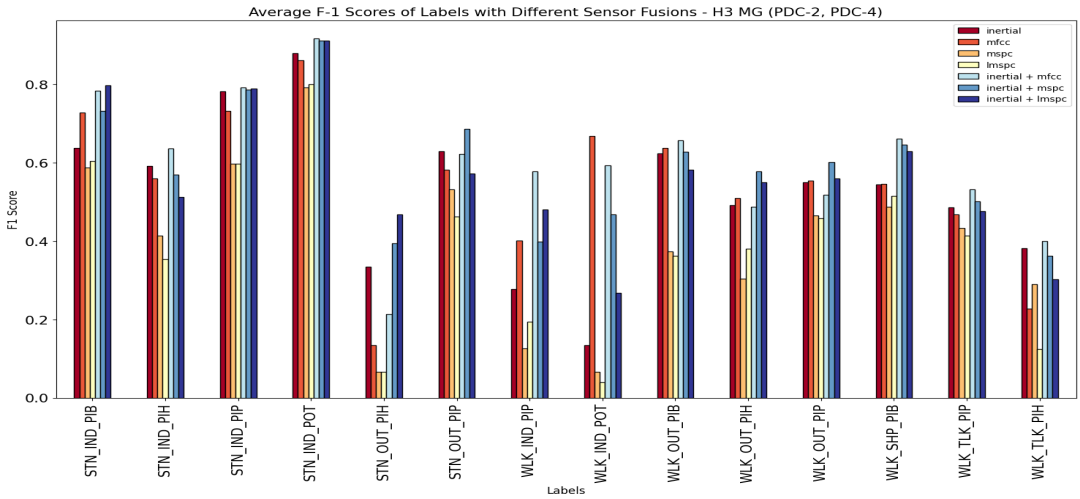
Test Group	Sensor	ACC	REC	SPE	BACC	PRE	Mac-F1	Mic-F1	W-F1
PDC-1	Inertial	0.986	0.807	0.953	0.880	0.811	0.800	0.986	0.987
	MFCC	0.986	0.787	0.962	0.874	0.795	0.780	0.986	0.986
	MSPC	0.973	0.544	0.890	0.717	0.601	0.560	0.973	0.974
	LMSPC	0.974	0.541	0.884	0.712	0.668	0.575	0.974	0.974
	Inertial & MFCC	0.986	0.824	0.962	0.893	0.792	0.803	0.986	0.987
	Inertial & MSPC	0.985	0.809	0.970	0.899	0.804	0.806	0.985	0.986
	Inertial & LMSPC	0.982	0.755	0.937	0.846	0.749	0.740	0.982	0.982
PDC-2	Inertial	0.805	0.631	0.949	0.790	0.670	0.642	0.805	0.805
	MFCC	0.781	0.592	0.942	0.767	0.620	0.599	0.781	0.783
	MSPC	0.683	0.495	0.918	0.707	0.496	0.488	0.683	0.683
	LMSPC	0.687	0.485	0.918	0.701	0.491	0.470	0.687	0.687
	Inertial & MFCC	0.847	0.684	0.971	0.827	0.677	0.670	0.847	0.848
	Inertial & MSPC	0.829	0.681	0.958	0.820	0.701	0.680	0.829	0.829
	Inertial & LMSPC	0.830	0.662	0.958	0.810	0.720	0.675	0.830	0.831
PDC-3	Inertial	0.781	0.591	0.969	0.780	0.590	0.578	0.781	0.782
	MFCC	0.804	0.629	0.972	0.800	0.598	0.598	0.804	0.805
	MSPC	0.815	0.616	0.973	0.795	0.613	0.603	0.815	0.815
	LMSPC	0.796	0.631	0.971	0.801	0.577	0.579	0.796	0.796
	Inertial & MFCC	0.867	0.700	0.981	0.840	0.708	0.689	0.867	0.867
	Inertial & MSPC	0.851	0.659	0.979	0.819	0.656	0.646	0.851	0.852
	Inertial & LMSPC	0.850	0.664	0.978	0.821	0.688	0.667	0.850	0.852
PDC-4	Inertial	0.510	0.470	0.925	0.697	0.446	0.437	0.510	0.511
	MFCC	0.506	0.521	0.923	0.722	0.519	0.501	0.506	0.506
	MSPC	0.399	0.331	0.906	0.619	0.345	0.319	0.399	0.400
	LMSPC	0.390	0.307	0.904	0.606	0.339	0.306	0.390	0.389
	Inertial & MFCC	0.564	0.650	0.940	0.782	0.690	0.635	0.564	0.564
	Inertial & MSPC	0.601	0.579	0.943	0.757	0.643	0.609	0.601	0.602
	Inertial & LMSPC	0.525	0.498	0.927	0.712	0.518	0.482	0.525	0.525

The first test group, *PDC-1* indicates, "Lying down" test group dependent to placement. In this group, even though inertial sensors only are very informative while distinguishing the context and phone placement, still audio and inertial sensor fusion increases the overall performance nearly 2%. The average score is 90% in inertial and MSPC feature fusion, despite the fact that the data is highly imbalanced.

In *PDC-2*, the model tries to predict both context and phone placement of main activity "Standing" at the same time. By using audio feature MFCC and inertial sensors in a combination, overall performance is measured as 83%. The MFCC and inertial combination also works well for *PDC-3*, for classes of "Sitting" activity. Confusion matrices are in figure 4.18 for fold id 3. For *PDC-4*, even though audio helps to improve overall performance up to 78%, still this score is below compared to other test groups. Figure 4.17 demonstrates the model performance over predicting each issued class in F1 measurement unit.

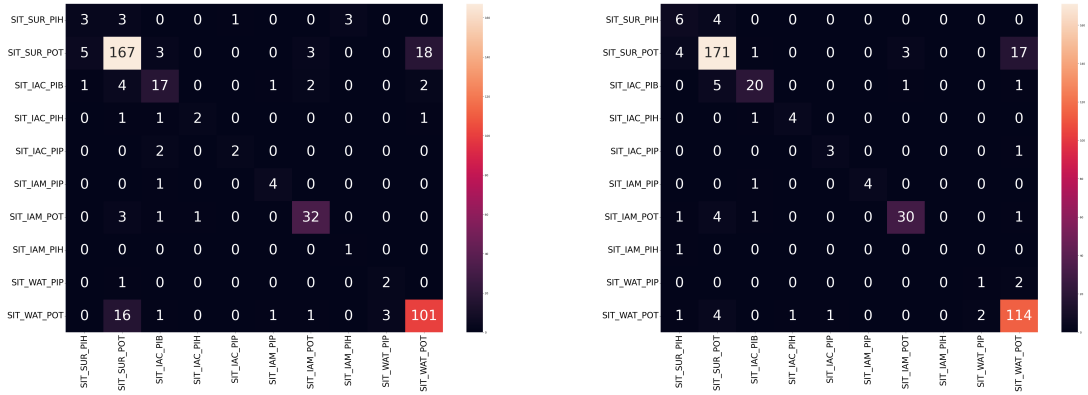


(a)



(b)

Figure 4.17: PDC Main Group Dependent F1 scores per context-placement - Holdout



(a) (b)

(c) (d)

(e) (f)

Figure 4.18: H3 Test Group - Placement Dependent Context-3 Confusion Matrices, Fold 3 (a-b) MFCC, Inertial & MFCC (c-d) MSPC, Inertial&MSPC, (e-f) LMSPC, Inertial&LMSPC, - 5-Fold

4.3 Discussion

The data set has been used in a variety of studies due to the extensive list of activities that may be used to replicate daily life that it represents, as well as the sufficient number of gathered samples to use timestamped data for activity recognition studies, since October 2017. Because the data collection contains asymmetrical real-life data, the majority of the researches report results with balanced accuracy. Table 4.18 lists studies done with this real-world dataset and proposed model performance comparison to others.

First attempt over the data set for behavioral context recognition with machine learning algorithms is reported by Vaizman et al [25] with Logistic Regression which they test single-sensor and sensor-fusion classifiers using 25 selected activity list. That work gives results overall Balanced Accuracy (BA) 71.8%, and they observed that using multiple sensor data gives better result than only single one. In the next work of Vaizman et al [52], they use multiple Multi-Layer-Perceptron (MLP) techique over 51 labels by adding 26 other activities which they did not consider for their previous work. Best performance is obtained with two-hidden-level with dimensions (16, 16), where BA is 77.3% average value of 51 label performance. Generally, accelerometer and gyroscope sensor data are used to predict body movements. In most of the works used the data set, rather than using all of the context labels, researches are focused on basic daily life activities such as "walking", "lying down" etc. Tarafdar et al. [53] uses only four of the primary physical activities and apply undersampling method to overcome imbalanced data issue. First, they conduct feature engineering, find out most useful features for each sensor and use them only, while they divide data as 60% train, 40% test. Cruciani et al. [55] also used undersampling to overcome unbalanced data issue among their activity list to work with, which includes five primary physical activities. They use inertial measurement unit (IMU) data. Average F1 score is reported as %52.7, with used CNN. In contrast most of the work in this dataset, they used audio-based modalities to predict environment of the activities. To do so, they create three groups named "Indoors", "Outdoors" and "Vehicle" based on original activities. Their F1 score over audio-based sensor modality is 21%. They do not conduct any fusion of audio and inertial sensors. Adaimi et al. [57] used 22 activities represented

in the data set. They conduct experiments with almost all sensor-modalities in the set, accelerometer, gyroscope, watch accelerometer, location, audio, phone state. To overcome imbalanced data, they use balanced class-weights. Furthermore, they applied active learning, both pull-based and stream-based approach to select data. They list results of each sensor modality used as feature, but early fusion generally serves best results, over others with reported balanced accuracy 79%. Since prediction over very highly correlated activity data is very hard, some researchs are focused on generating their own activity approaches by combining multi-daily activities as one activity and conduct classification over new contexts. For example, Ehatisham-Ul-Haq et al [54] firstly use six of the primary activities, where other activities are dependent to any one of them such as "lying when sleeping with phone on table". They define 29 fine-grained activities by doing so. Furthermore, they apply feature reduction technique CfsSubetSel to find and use features from sensors only gives best distinction performance from all features. By doing so, they use only 12 features out of 26 per each sensor. Even though they do experiments with other classifiers as well, RF gives best result for six physical activities balanced accuracy 83%, 29 context-dependent activities 86%. Grouping activities to issue with less context approach is applied by others as well like Asim et al [56]. They define activities of daily life as primary and secondary. They named same primary activities in Ehatisham-Ul-Haq et al [54] are used while they generate 15 secondary contexts which are combination of activities, for example, "Standing and Shopping". They also applied feature reduction during preparing the data. They use only 18 most informative features from phone accelerometer for their work. The best results for accuracy are obtained again with RF where balanced accuracy is reported as 80% for six primary activities while it is 77% for secondary contexts. Ehatisham-ul-Haq et al [58] represent another work with same approach as their previous work, Ehatisham-ul-Haq et al [54]. As difference from previous work, they level activities as physical activities, behavioral activities and phone positions. They apply three level based testing while also within the same level, groups are divided with respect to physical activity. Furthermore, they conduct feature engineering and decide to work with 18 extracted over 26 features from smartphone and smartwatch accelerometer sensor data. Two-level classifier is applied, activity and context. With RF, balanced accuracy score is reported as 80% for 5 main activities while it is 76% second level, 13 classes. For phone position

dependent ones, results especially for "Sitting" and "Standing" main groups are very high, in overall 83% balanced accuracy score is reported.

In this study, the activities are arranged hierarchically to avoid confusing the proposed model in circumstances when multiple activities and contexts are done at same time, similar to [54], [56] and [58]. Furthermore, test are conducted in two different cross-validation method; holdout and 5-fold.

In model evaluation, there are two main approaches, first, user-defined labels are tested within the same hierarchy level, without grouping based on primary physical activity. With the audio fusion effect with inertial sensors widely used in this field, promising results are obtained considering real-world structure and imbalanced data. For holdout validation, in case of primary activities, balanced accuracy score is measured as 91% for classification done with five physical activities, average from five iterations, which outperforms all of the previously reported results by a sufficiently big margin, scores obtained using only five physical primary activity is reported as 88%([53]) and 80%([54]), even though those works use high number of features compared to this work. The average balanced accuracy measured with 5-fold cross validation is 90% for five class case, which still outperforms existing studies. For six primary physical activity class, activity "bicycling" is added into the system. The average balanced accuracy score is measured as 90% which exceeds all other researches utilizing those classes during model evaluation, for holdout. Similarly for 5-fold, model performs with 91% performance while distinguishing six primary physical activities.

In comparison to the preceding layer, the multiclass classification process becomes more challenging at hierarchy level two because more complex activity-context pairs with variable combinations are regarded as classes. For this level, there are two main types of testing, one is main group independent and another is main group dependent. In case of main group independent testing, there are 15 fine-grained activity classes constructed by multiple contexts and activities. Proposed ligthweight model outperforms existing study [56] by 2% higher rate with the same number of classes with overall balanced accuracy score measured 79%, for holdout cross validation. In case of 5-fold cross validation, the overall score is measured as 81%, which is 2%

higher from the average holdout result, where 4% higher than [56]. This is achieved using inertial and audio sensor fusion since only inertial data does not perform as well as combination. In case of main group dependent case, there are four main groups, which yield different scores. In case of level two testing, for holdout cross validation, overall balanced accuracy of four distinct test group is 82%, which exceeds highest reported BA score 76% with the exact same label configuration in [58]. Similarly, 83% balanced accuracy score is obtained in the same hierarchical configuration, which is significantly, around 7%, greater than [58]. Comparing by test case, only placement independent activities of group "walking" is lower than their work while other three group performance is extremely higher than reported scores for both of the validation methods.

The depth of tree is end with level three, where smartphone placement become crucial element while constructing labels with activity-context pair information inherited from previous layer. Due to framing approach and lack of labelled information from users, there are no "running, exercise, phone on pocket" for any of the 60 users in the pre-processed data. Therefore, hierarchy level three does not include that user-defined context. In research [54], researchers use three sensor modalities with multiple features selected in data processing case and obtain 86% in level three case, main group independent test grouping. Even though results at the same configuration in this study does not exceed this score for both of the cross validation methods, still audio data boost overall system performance when used with inertial sensors. In case of main group dependent test case, however, proposed solution surpasses state of art solutions, with a sufficient margin. Average score based on iteration mean results of four activity dependent groups become 85%, which is nearly 2% higher than reported results of [58], which is 83%, for holdout validation. The scores are nearly 1.5% higher from existing study [58] in case of 5-fold cross validation, in level three, main group dependent testing. Their two test groups, "sitting" and "standing" derivatives, perform with very high scores while remaining two is reported extremely low compared to those. In this study, however, results for three groups are very close, but in case of "walking", result is slightly lower.

Table 4.18: Studies using the real-world ExtraSensory dataset and results

Study	#Act.	Activity Type	Sensors	Evaluation	Classifier	Metric	Result
[25]	25	Behavioral Context	PA, WA, L, G, PS, AU	5-Fold	LR	BA	0.72
[52]	51	Behavioral Context	PA, WA, L, G, PS, AU	5-Fold	MLP	BA	0.76
[53]	5	Physical Activity	PA, WA, G, M	Holdout	AdaBoost	A	0.88
[55]	5 3	Physical Activity Location	PA, G AU	5-Fold	CNN	F1	0.52 0.21
[57]	22	Behavioral Context	PA, WA, L G, PS, AU	Holdout	LR	BA	0.79
[54]	6 29	Physical Activity Context with PP	PA, WA, G	5-Fold	RF	BA	0.83 0.86
[56]	6 15	Physical Activity Behavioral Context	PA	10-Fold	RF	BA	0.80 0.77
[58]	5 13 29	Physical Activity Behavioral Context Context with PP	PA, WA	5-Fold	RF	BA	0.80 0.76 0.83
Proposed (MGI)	**6** **15** **28**	**Physical Activity** **Behavioral Context** **Context with PP**	**PA, WA, G, M, AU**	**Holdout**	**CNN**	**BA**	**0.90** **0.79** **0.77**
Proposed (MGD)	**5** **13** **28**	**Physical Activity** **Behavioral Context** **Context with PP**	**PA, WA, G, M, AU**	**Holdout**	**CNN**	**BA**	**0.91** **0.82** **0.85**
Proposed (MGI)	**6** **15** **28**	**Physical Activity** **Behavioral Context** **Context with PP**	**PA, WA, G, M, AU**	**5-Fold**	**CNN**	**BA**	**0.91** **0.81** **0.76**
Proposed (MGD)	**5** **13** **28**	**Physical Activity** **Behavioral Context** **Context with PP**	**PA, WA, G, M, AU**	**5-Fold**	**CNN**	**BA**	**0.90** **0.83** **0.84**

*A:Accuracy, BA:Balanced Accuracy, AU:Audio, PS:Phone State, G:Gyroscope, M:Magnetometer, PA:Phone Accelerometer, WA:Watch Accelerometer, L:Location, MGI: Main Group Independent, MGD: Main Group Dependent, PP:Phone Placement

The proposed solution performs with sufficient and promising results even with the data of various distinct characterized people operating not based on pre-defined script but based on their regular habits. The time window is set at 20 rpw to express each activity flow better while keeping sample size maximum in this highly imbalanced dataset. Defining multiple activities tend to occur simultaneously in a form with a concatenation and considering those as one class, with respect to hierarchy adjust model to be more applicable in real-world HAR task. Inertial sensors, not required wearing some other measurement devices but embedded inside smartphone and smartwatch give information about what does the primary activity of the user at any time. Audio data, even though is sliced to little window size unlike general usage, illustrates highly effective results both in its own at some cases and in a fusion with inertial sensory information. Both MFCC and MSPC performances are good while at some cases, MFCC suprass MSPC and vice versa. Furthermore, proposed feature processing emphasis the importance of temporal change during image generation and activity recognition.

In the thesis, a lightweight deep learning model is utilized to perform recognition task. Model occupies 1.9 MB to 6.2 MB memory space based on sample number of activities involved and size of the data, which differ for inertial images to audio and fusion images, where each of them have 20 pixel width and all are 3 channelled. Table 4.19 demonstrates average memory space of the model as well as input sizes.

Table 4.19: Average Memory Space of Model

Input Type	#Height(P)	Single Image Size	#Params(Model)	Memory Space(Model)
Inertial	4	102 B	228,586	1.9 MB
MFCC	14	613 B	392,426	3.2 MB
MSPC	33	1.0 KB	703,722	5.7 MB
Inertial & MFCC	18	667 B	457,962	3.7 MB
Inertial & MSPC	37	1.1 KB	769,258	6.2 MB

*B:Bytes, KB:Kilobytes, MB:Megabytes, P:Pixel

After testing phase, models that work with different types of images are trained with whole chunk of the data of 60 users, to form create pre-trained model, for future use. It could be inferred intuitively that when the pre-trained version of this lightweight

model is integrated into any user's phone, it could perform automatic activity recognition by operating in the background process for extremely brief periods of time, with using embedded sensor data. From the Table 4.19 it is concluded that model occupy limited space. Furthermore, from the analysis conducted during test phase, overall classification prediction time of the model for single image is measured as $5x10^{-4}$ seconds, measured with a virtual machine which has Ubuntu 20.04. Based on these findings, it could be concluded that if the model is integrated into any user's phone, it will take up little space in the phone's memory while working in the background not effecting phone performance much while conducting real-time activity recognition automatically.

CHAPTER 5

CONCLUSION

The most challenging part of real-life applications is that each person uses theirs own device, smartphone and smartwatch in this case, and perform natural behavior on their own environment, without someone telling what to do in which duration. Even though scripted scenarios are favored for activity recognition tasks because they provide fewer problems and produce better results, with the help of developing technology, more accurate findings are obtained by using models that could grasp the activities and contexts of bigger groups of people without imposing additional burdens on them.

The processing time of an activity, therefore recording the sensory information is completely user dependent in real-world case. Therefore, optimal solution could be utilized to operate with user-independent models. Data analysis and preprocessing is important in that aspect. The promising success of the idea of recreating the sensor information in image format is riveted with this work, for both inertial sensors and audio sensor. Simulating activities occurring over a period of time in their corresponding pattern with the proposed differential data representation greatly improves the model's performance. To detect an activity with sensor data, it takes a particular length of time and repetition, which varies depending on the sensor. Therefore, window size selection is very important. Even though recognition with audio requires wider window size, to maximize sample size contributing in the processed data, audio is used with little window size. Data representation technique is very crucial to work with different systems. Since in this thesis, a performance is illustrated with a lightweight deep learning model, image format is chosen to work with a CNN model. In such system with there is a large bias between samples of activities, proposed

model that utilize data representation performs outstanding results. The difficulty of multi-label classification has been solved by using a hierarchy to define the activity, context, and placement with each hierarchy. At the same time, people could engage in an activity and engage in a context. Furthermore, since device-placement is leaved up to user preference, recognition become possible with this approach regardless of user habits.

REFERENCES

[1] Michael Bironneau, & Toby Coleman. (2019). Machine Learning with Go Quick Start Guide: Hands-on Techniques for Building Supervised and Unsupervised Machine Learning Workflows. Packt Publishing.

[2] Hinton, Geoffrey; Sejnowski, Terrence (1999). Unsupervised Learning: Foundations of Neural Computation. MIT Press. ISBN 978-0262581684.

[3] Olivier Chapelle, Bernhard Schölkopf, & Alexander Zien. (2006). Semi-Supervised Learning. The MIT Press.

[4] Ethem Alpaydin. (2016). Machine Learning: The New AI. The MIT Press.

[5] Richard S. Sutton, & Andrew G. Barto. (1998). Reinforcement Learning: An Introduction. A Bradford Book.

[6] Cortes, C., Vapnik, V. Support-vector networks. Mach Learn 20, 273–297 (1995). https://doi.org/10.1007/BF00994018

[7] Kwon, S. J. (2011). Artificial Neural Networks. Nova Science Publishers, Inc.

[8] Ding, B., Qian, H., & Zhou, J. (2018). Activation functions and their characteristics in deep neural networks. 2018 Chinese Control And Decision Conference (CCDC), Chinese Control And Decision Conference (CCDC), 2018, 1836–1841. https://doi.org/10.1109/CCDC.2018.8407425

[9] Schmidhuber, J. (2014). Deep Learning in Neural Networks: An Overview. https://doi.org/10.1016/j.neunet.2014.09.003

[10] Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., & Lew, M. S. (2016). Deep learning for visual understanding: A review. Neurocomputing, 187, 27–48. https://doi.org/10.1016/j.neucom.2015.09.116

[11] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Proceedings of the NIPS, 2012.

[12] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: deep neural networks with multitask learning. In Proceedings of the 25th international conference on Machine learning, pages 160–167. ACM, 2008.

[13] Dipanjan Sarkar, Raghav Bali, & Tamoghna Ghosh. (2018). Hands-On Transfer Learning with Python: Implement Advanced Deep Learning and Neural Network Models Using TensorFlow and Keras. Packt Publishing.

[14] Ian Goodfellow, Yoshua Bengio, & Aaron Courville. (2016). Deep Learning. The MIT Press.

[15] Anurag Bhardwaj, Wei Di, & Jianing Wei. (2018). Deep Learning Essentials: Your Hands-on Guide to the Fundamentals of Deep Learning and Neural Network Modeling. Packt Publishing.

[16] Kim P. (2017) Convolutional Neural Network. In: MATLAB Deep Learning. Apress, Berkeley, CA. https://doi.org/10.1007/978-1-4842-2845-6_6

[17] Sepp Hochreiter, Jürgen Schmidhuber; Long Short-Term Memory. Neural Comput 1997; 9 (8): 1735–1780. doi: https://doi.org/10.1162/neco.1997.9.8.1735

[18] Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation.

[19] Wang, J.; Chen, Y.; Hao, S.; Peng, X.; Hu, L. Deep Learning for Sensor-based Activity Recognition: A Survey. Comput. Vis. Pattern Recognit. 2017, 119, 3–11.

[20] Y. Chen and C. Shen, "Performance Analysis of Smartphone-Sensor Behavior for Human Activity Recognition," in IEEE Access, vol. 5, pp. 3095-3110, 2017, doi: 10.1109/ACCESS.2017.2676168.

[21] Ricardo Chavarriaga, Hesam Sagha, Alberto Calatroni, Sundara Tejaswi Digumarti, Gerhard Tröster, José del R. Millán, Daniel Roggen, The Opportunity challenge: A benchmark database for on-body sensor-based activity recognition, Pattern Recognition Letters, Volume 34, Issue 15, 2013, Pages 2033-2042, ISSN 0167-8655, https://doi.org/10.1016/j.patrec.2012.12.014.

[22] R. Chavarriaga et al., "The opportunity challenge: A benchmark database for on-body sensor-based activity recognition," Pattern Recognit. Lett., vol. 34, no. 15, pp. 2033–2042, Nov. 2013, doi: 10.1016/j.patrec.2012.12.014

[23] Anguita, D.; Ghio, A.; Oneto, L.; Parra, X.; Reyes-Ortiz, J.L. A Public Domain Dataset for Human Activity Recognition Using Smartphones. In Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges, Belgium, 24–26 April 2013; pp. 24–26.

[24] 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. In Proceedings of the 6th International Work-Conference on Ambient Assisted Living an Active Ageing (IWAAL 2014), Belfast, UK,2–5 December 2014.

[25] Vaizman, Y., Ellis, K., and Lanckriet, G. "Recognizing Detailed Human Context In-the-Wild from Smartphones and Smartwatches". IEEE Pervasive Computing, vol. 16, no. 4, October-December 2017, pp. 62-74. doi:10.1109/MPRV.2017.3971131

[26] Zhang, M.; Sawchuk, A.A. USC-HAD: A Daily Activity Dataset for Ubiquitous Activity Recognition Using Wearable Sensors. In Proceedings of the 2012 ACM Conference on Ubiquitous Computing, Pittsburgh, PA, USA, 5–8 September 2012; p. 1036.

[27] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," ACM SIGKDD Explorations Newslett., vol. 12, no. 2, pp. 74–82, Mar. 2011.

[28] Zheng Y., Liu Q., Chen E., Ge Y., Zhao J.L. (2014) Time Series Classification

Using Multi-Channels Deep Convolutional Neural Networks. In: Li F., Li G., Hwang S., Yao B., Zhang Z. (eds) Web-Age Information Management. WAIM 2014. Lecture Notes in Computer Science, vol 8485. Springer, Cham. https://doi.org/10.1007/978-3-319-08010-9_33

[29] O. Baños, M. Damas, H. Pomares, I. Rojas, M. A. Tóth, and O. Amft, "A benchmark dataset to evaluate sensor displacement in activity recognition," in Proc. ACM Conf. Ubiquitous Comput., 2012, pp. 1026–1035

[30] K. Altun, B. Barshan, and O. Tunçel, "Comparative study on classifying human activities with miniature inertial and magnetic sensors," Pattern Recognit., vol. 43, no. 10, pp. 3605–3620, Oct. 2010

[31] A. Stisen, H. Blunck, S. Bhattacharya, T. S. Prentow, M. B. Kjærgaard, A. Dey, T. Sonne, and M. M. Jensen, "Smart devices are different: Assessing and mitigatingmobile sensing heterogeneities for activity recognition," in Proc. 13th ACM Conf. Embedded Netw. Sensor Syst., 2015, pp. 127–140.

[32] Lima, W.S., Souto, E.J., El-Khatib, K., Jalali, R., & Gama, J. (2019). Human Activity Recognition Using Inertial Sensors in a Smartphone: An Overview. Sensors (Basel, Switzerland), 19.

[33] McFee, Brian, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. "librosa: Audio and music signal analysis in python." In Proceedings of the 14th python in science conference, pp. 18-25. 2015.

[34] Ferhat Attal, Samer Mohammed, Mariam Dedabrishvili, Faicel Chamroukhi, Latifa Oukhellou, Yacine Amirat. (2015). Physical Human Activity Recognition Using Wearable Sensors. Sensors, 15(12), 31314–31338. https://doi.org/10.3390/s151229858

[35] M. Mario, "Human Activity Recognition Based on Single Sensor Square HV Acceleration Images and Convolutional Neural Networks," in IEEE Sensors Journal, vol. 19, no. 4, pp. 1487-1498, 15 Feb.15, 2019, doi: 10.1109/JSEN.2018.2882943.

[36] Andrey Ignatov, Real-time human activity recognition from accelerometer data using Convolutional Neural Networks, Applied Soft Computing, Volume 62, 2018, Pages 915-922, ISSN 1568-4946, https://doi.org/10.1016/j.asoc.2017.09.027.

[37] Mohammad Mehedi Hassan, Shamsul Huda, Md Zia Uddin, Ahmad Almogren, and Majed Alrubaian. 2018. Human Activity Recognition from Body Sensor Data using Deep Learning. J. Med. Syst. 42, 6 (June 2018), 1–8. DOI:https://doi.org/10.1007/s10916-018-0948-z

[38] W. Gao, L. Zhang, W. Huang, F. Min, J. He and A. Song, "Deep Neural Networks for Sensor-Based Human Activity Recognition Using Selective Kernel Convolution," in IEEE Transactions on Instrumentation and Measurement, vol. 70, pp. 1-13, 2021, Art no. 2512313, doi: 10.1109/TIM.2021.3102735.

[39] Minhyuk Jung, Seokho Chi, Human activity classification based on sound recognition and residual convolutional neural network, Automation in Construction, Volume 114, 2020, 103177, ISSN 0926-5805, https://doi.org/10.1016/j.autcon.2020.103177.

[40] Q. Teng, K. Wang, L. Zhang and J. He, "The Layer-Wise Training Convolutional Neural Networks Using Local Loss for Sensor-Based Human Activity Recognition," in IEEE Sensors Journal, vol. 20, no. 13, pp. 7265-7274, 1 July1, 2020, doi: 10.1109/JSEN.2020.2978772.

[41] Y. Wang, S. Cang and H. Yu, "A Data Fusion-Based Hybrid Sensory System for Older People's Daily Activity and Daily Routine Recognition," in IEEE Sensors Journal, vol. 18, no. 16, pp. 6874-6888, 15 Aug.15, 2018, doi: 10.1109/JSEN.2018.2833745.

[42] J. Huang, S. Lin, N. Wang, G. Dai, Y. Xie and J. Zhou, "TSE-CNN: A Two-Stage End-to-End CNN for Human Activity Recognition," in IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 1, pp. 292-299, Jan. 2020, doi: 10.1109/JBHI.2019.2909688.

[43] J. A. Stork, L. Spinello, J. Silva and K. O. Arras, "Audio-based human activity recognition using Non-Markovian Ensemble Voting," 2012 IEEE

RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication, 2012, pp. 509-514, doi: 10.1109/ROMAN.2012.6343802

[44] C. Xu, D. Chai, J. He, X. Zhang and S. Duan, "InnoHAR: A Deep Neural Network for Complex Human Activity Recognition," in IEEE Access, vol. 7, pp. 9893-9902, 2019, doi: 10.1109/ACCESS.2018.2890675.

[45] Z. Chen, Le Zhang, Z. Cao and J. Guo, "Distilling the Knowledge From Handcrafted Features for Human Activity Recognition," in IEEE Transactions on Industrial Informatics, vol. 14, no. 10, pp. 4334-4342, Oct. 2018, doi: 10.1109/TII.2018.2789925.

[46] Liangying Peng, Ling Chen, Zhenan Ye, and Yi Zhang. Aroma: A deep multitask learning based simple and complex human activity recognition method using wearable sensors. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2(2):1–16, 2018.

[47] Minh Dang, L., Min, K., Wang, H., Jalil Piran, M., Hee Lee, C., Moon, H. (2020). Sensor-based and vision-based human activity recognition: A comprehensive survey. Pattern Recognition, 108. https://doi.org/10.1016/j.patcog.2020.107561

[48] L. Roland , L. Lidauer , G. Sattlecker , F. Kickinger , W. Auer , V. Sturm , D. Efrosinin , M. Drillich , M. Iwersen , Monitoring drinking behavior in buck- et-fed dairy calves using an ear-attached tri-axial accelerometer: a pilot study, Comput. Electron. Agric. 145 (2018) 298–301.

[49] Martínez-Pérez, F.E.; González-Fraga, J.A.; Cuevas-Tello, J.C.; Rodríguez, M.D. Activity Inference for Ambient Intelligence Through Handling Artifacts in a Healthcare Environment. Sensors 2012, 12, 1072–1099.

[50] Mitchell, E.; Monaghan, D.; O'Connor, N.E. Classification of sporting activities using smartphone accelerometers. Sensors 2013, 13, 5317–5337.

[51] H. Y. Yatbaz, E. Ever and A. Yazici, "Activity Recognition and Anomaly Detection in E-Health Applications Using Color-Coded Representation and

Lightweight CNN Architectures," in IEEE Sensors Journal, vol. 21, no. 13, pp. 14191-14202, 1 July1, 2021, doi: 10.1109/JSEN.2021.3061458.

[52] Vaizman, Y., Weibel, N., and Lanckriet, G. "Context Recognition In-the-Wild: Unified Model for Multi-Modal Sensors and Multi-Label Classification". Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT), vol. 1, no. 4. December 2017. doi:10.1145/3161192

[53] Tarafdar, P., Bose, I. (2021). Recognition of human activities for wellness management using a smartphone and a smartwatch: A boosting approach. Decision Support Systems, 140, N.PAG. https://doi.org/10.1016/j.dss.2020.113426

[54] Ehatisham-Ul-Haq, M. (1), Azam, M. A. (1,2), Amin, Y. (1), Naeem, U. (3). (n.d.). C2FHAR: Coarse-to-Fine Human Activity Recognition with Behavioral Context Modeling Using Smart Inertial Sensors. IEEE Access, 8, 7731–7747. https://doi.org/10.1109/ACCESS.2020.2964237

[55] Cruciani, F., Vafeiadis, A., Nugent, C., Cleland, I., McCullagh, P., Votis, K., Giakoumis, D., Tzovaras, D., Chen, L., Hamzaoui, R. (2020). Feature learning for Human Activity Recognition using Convolutional Neural Networks: A case study for Inertial Measurement Unit and audio data. CCF Transactions on Pervasive Computing and Interaction, 2(1), 18. https://doi.org/10.1007/s42486-020-00026-2

[56] Asim, Y. (1), Azam, M. A. (1), Ehatisham-Ul-Haq, M. (1), Naeem, U. (2), Khalid, A. (3). (n.d.). Context-Aware Human Activity Recognition (CAHAR) in-the-Wild Using Smartphone Accelerometer. IEEE Sensors Journal, 20(8), 4361–4371. https://doi.org/10.1109/JSEN.2020.2964278

[57] Adaimi, R., Thomaz, E. (2019). Leveraging Active Learning and Conditional Mutual Information to Minimize Data Annotation in Human Activity Recognition. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 3(3), 1–23. https://doi.org/10.1145/3351228

[58] Ehatisham-ul-Haq, M., Azam, M. A. (2020). Opportunistic sensing for inferring in-the-wild human contexts based on activity pattern recognition using smart computing. Future Generation Computer Systems, 106, 374–392. https://doi.org/10.1016/j.future.2020.01.003

[59] Pushpajit Khaire, Praveen Kumar, Javed Imran, Combining CNN streams of RGB-D and skeletal data for human activity recognition, Pattern Recognition Letters, Volume 115, 2018, Pages 107-116, ISSN 0167-8655, https://doi.org/10.1016/j.patrec.2018.04.035.

[60] Li, W., Wen, L., Chang, M. C., Nam Lim, S., Lyu, S. (2017). Adaptive RNN tree for large-scale human action recognition. In Proceedings of the IEEE International Conference on Computer Vision (pp. 1444-1452).

[61] C. Jia, M. Shao, S. Li, H. Zhao and Y. Fu, "Stacked Denoising Tensor Auto-Encoder for Action Recognition With Spatiotemporal Corruptions," in IEEE Transactions on Image Processing, vol. 27, no. 4, pp. 1878-1887, April 2018, doi: 10.1109/TIP.2017.2781299.

[62] A. Prati , C. Shan , K.I.-K. Wang , Sensors, vision and networks: from video surveillance to activity recognition and health monitoring, J. Ambient Intell. Smart Environ. 11 (1) (2019) 5–22 .

[63] K.S. Kumar , R. Bhavani , Human activity recognition in egocentric video using hog, gist and color features, Multimed. Tools Appl. (2018) 1–17 .

[64] P.K. Roy , H. Om , Suspicious and violent activity detection of humans using hog features and SVM classifier in surveillance videos, in: Advances in Soft Computing and Machine Learning in Image Processing, Springer, 2018, pp. 277–294 .

[65] A. Thyagarajmurthy , M. Ninad , B. Rakesh , S. Niranjan , B. Manvi ,Anomaly de- tection in surveillance video using pose estimation, in: Emerging Research in Electronics, Computer Science and Technology, Springer, 2019, pp. 753–766.

[66] L. Martínez-Villaseñor , H. Ponce , A concise review on sensor signal acquisition and transformation applied to human activity recognition and hu-

man–robot interaction, International Journal of Distributed Sensor Networks 15 (6) (2019) . 1550147719853987

[67] Tsitsoulis, A., Bourbakis, N. (2013). A first stage comparative survey on human activity recognition methodologies. International Journal on Artificial Intelligence Tools, 22(06), 1350030.

[68] A. Ullah, K. Muhammad, J. Del Ser, S. W. Baik and V. H. C. de Albuquerque, "Activity Recognition Using Temporal Optical Flow Convolutional Features and Multilayer LSTM," in IEEE Transactions on Industrial Electronics, vol. 66, no. 12, pp. 9692-9702, Dec. 2019, doi: 10.1109/TIE.2018.2881943.

[69] Ahmad Jalal, Yeon-Ho Kim, Yong-Joong Kim, Shaharyar Kamal, Daijin Kim, Robust human activity recognition from depth video using spatiotemporal multi-fused features, Pattern Recognition, Volume 61, 2017, Pages 295-308, ISSN 0031-3203, https://doi.org/10.1016/j.patcog.2016.08.003.

[70] Wang, P., Wang, S., Gao, Z., Hou, Y., Li, W. (2017). Structured images for RGB-D action recognition. In Proceedings of the IEEE International Conference on Computer Vision Workshops (pp. 1005-1014).

[71] Yong Du, W. Wang and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1110-1118, doi: 10.1109/CVPR.2015.7298714.

[72] Chollet, F., others. (2015). Keras. GitHub. Retrieved from https://github.com/fchollet/keras

[73] Imen Jegham, Anouar Ben Khalifa, Ihsen Alouani, Mohamed Ali Mahjoub, Vision-based human action recognition: An overview and real world challenges, Forensic Science International: Digital Investigation, Volume 32, 2020, 200901, ISSN 2666-2817, https://doi.org/10.1016/j.fsidi.2019.200901.

[74] Rahmani, H., Mian, A., Shah, M. (2017). Learning a deep model for human action recognition from novel viewpoints. IEEE transactions on pattern analysis and machine intelligence, 40(3), 667-681.