Assessing the Robustness of HAR Deep Learning Models against Variability

Azhar Ali Khaked

A Thesis

in

The Department

of

Electrical and Computer Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of Master of Applied Science at

Concordia University

Montreal, Quebec, Canada

November 2024

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Azhar Ali Khaked**

Entitled: **Assessing the Robustness of HAR Deep Learning Models against Variability**

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science (Electrical and Computer Engineering)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

_____ Chair

Habib Benali

_____ Examiner

Yan Liu

_____ Supervisor

Paula Lago

Approved by _____

Jun Cai

Graduate Program Director

25 November 2024 _____

Mourad Debbabi

Faculty of Engineering and Computer Science

Abstract

Assessing the Robustness of HAR Deep Learning Models against Variability

Azhar Ali Khaked

Deep learning (DL) Human Activity Recognition (HAR) models using wearable inertial measurement unit (IMU) sensors have shown great promise in applications like continuous healthcare monitoring and early disease prediction. However, most DL HAR models remain untested in real-world scenarios laden with variabilities; rather, they are trained and tested on constrained and closely curated HAR datasets that assume an ideal setting. This thesis explains the effects of real-world variabilities like subject, device, position, and orientation on the performance of DL HAR models. Due to the inability of existing datasets to isolate variabilities, we collect our own, the HARVAR dataset. We isolated the effect of different variabilities and provided a nuanced understanding of how each affects DL HAR models' performance. Maximum Mean Discrepancy (MMD) was used to quantify shifts in data distribution due to each isolated variability and drew a relationship between the drop in performance and the change in data distribution. The REALDISP dataset was used to perform a case study to understand the effects of compounded and unisolated variabilities in the real world. This study found that different variabilities have varying effects on the DL HAR model performance, from insignificant to detrimental. We showed a negative correlation between the MMD and the performance drop of the DL HAR models in the results drawn from both HARVAR and REALDISP datasets. The study emphasizes the need for more robust models and the development of pre-processing methodologies to optimize the IMU data for training robust DL HAR models.

Table of Contents

1 Introduction **1**

1.1 Objectives . . . 3

1.2 Thesis Structure . . . 3

1.3 Publications . . . 4

2 Related Work **5**

2.1 Deep Learning Models in HAR . . . 5

2.2 Limitations of Current HAR Evaluation and Data Heterogeneity . . . 8

2.3 Measuring Distribution Shifts . . . 10

2.4 Variability in HAR . . . 10

3 HARVAR Dataset **13**

3.1 Existing Datasets and their limitations . . . 14

3.2 Sensors and Sensor Placement . . . 16

3.3 Activity Protocol for Data Collection . . . 17

3.4 Sensor Data Transmission, Storage, Syncing, and Labeling . . . 20

3.5 Data Overview . . . 21

3.6 Potential Use Cases of the dataset . . . 24

4 Method to evaluate the Effect of Variability on Deep Learning HAR Models **25**

4.1 Experimental Protocol to Evaluate Model Performance under Variability . . . 25

4.1.1 Model Training . . . 27

4.1.2 Model Performance Evaluation 30
4.1.3 Measuring Variability with MMD 30
4.2 Measuring Compounding Effects of Variability 32

5 Results of Variability Evaluation 34
5.1 Variability Impacts on Model Performance 34
5.1.1 Orientation Variability 35
5.1.2 Positional Variability 36
5.1.3 Device Variability 38
5.1.4 Subject Variability 38
5.2 Understanding Model Performance with MMD Metric 39
5.2.1 MMD to Explain Orientation, Position and Device Variability 40
5.2.2 MMD to Explain Subject Variability 45
5.2.3 MMD Correlation to F1 Score 45
5.3 Gender-Based Analysis using HARVAR Results 47
5.3.1 GBA+ of Orientation Variability on DL HAR Models 47
5.4 Compounding Variability Effects in Real-Life Scenarios (REALDISP Case Study) 50

6 Conclusion 54
6.1 Study Implications 54
6.1.1 Position and Orientation Variability Implications on Real-World Scenarios 54
6.1.2 Device Variability has a High Impact on DL Model Performance. 55
6.1.3 Subject Variability and the Need for Diverse Training Data 55
6.1.4 Larger MMD Correlate with Smaller F1-Score, with Limitations 56
6.1.5 No Significant Differences in Performance Change Across the Three Models 57
6.2 Study Limitations 57
6.3 Summary and Future Work 58

Chapter 1

Introduction

Many human activities involve repetitive or periodic physical movements, such as walking, running, and various exercises. Activities like cooking, eating, and cleaning are more complex and variable, often involving diverse movements comprising shorter recognizable motions. Inertial Measurement Unit (IMU) sensors can capture these movements as signals that can be used to recognize models and classify what activities are being performed. This classification process is called Human Activity Recognition (HAR) and can be used in healthcare monitoring[44], sports performance analysis[46], industrial and workplace safety[69], home automation[21], and gaming[35] and many other applications. IMU sensors can be found in smart wearable devices such as watches, earbuds, and glasses, allowing HAR to perform ubiquitously without special arrangements.

As HAR is a classification problem, HAR models can either be unsupervised or supervised, where unsupervised HAR models group similar activities together, and supervised HAR models learn from labeled sensor data to identify activities. Supervised HAR models can be broadly divided into traditional Machine Learning (ML) models or Deep Learning (DL) models. Conventional Machine Learning (ML) models, such as Support Vector Machines and Random Forests[4], have been applied to HAR but require manual feature engineering and domain expertise. In contrast, Deep Learning (DL) models such as Convolution Neural networks (CNNs) have been shown to automate feature extraction from input data in images [33] and sound [43]. This automation

has driven the development of various DL HAR models, showing promising results in activity classification from wearable IMU sensors.

This work focuses on DL-supervised learning models, where the training data must be pre-labeled or annotated to learn to classify signals into activities. Labeling HAR data is a laborious task as, unlike natural language or vision data, simply looking at the IMU data can not allow annotators to know what label to assign them. To overcome this, researchers usually use videos to allow annotators to label activity time stamps or assign an annotator to label activities as they are performed in a laboratory setting. Due to the amount of manual labor that goes into labeling IMU HAR data, the amount of labeled IMU HAR data is scarce. DL models require a vast quantity of data to perform, and despite the excellent performance depicted by existing DL HAR models, the lack of labeled sensor data is a big challenge for supervised HAR [86].

Due to the difficulties with HAR sensor data collection, most data collection is done in controlled lab settings with small participant groups. This setup limits participant diversity, making it harder for DL HAR models to generalize to real-world scenarios. Standard protocols also often require participants to perform activities in a prescribed manner, removing any movement variations between people based on habit or demographics. Lastly, researchers usually attach IMU sensors in ideal configurations during data collection, minimizing real-world variations due to wearing practices.

Real-world variations in IMU data occur due to device changes, wearing habits, and varying users. Any physical change to an IMU sensor that alters its measurements can be considered a variability source [15]. Device variability results from hardware variations, including sensor sensitivity, sampling rate, range, and noise [68]. Wearing variability arises from changes in the position or orientation of the sensor worn by the user [83]. Subject variability arises due to users performing actions in varying ways [32, 79].

Despite DL HAR models' high performance, they have been trained and tested on constrained datasets. Assessing DL HAR models on data with no variability provides a limited understanding of their performance and robustness. Variability induces a distribution shift in the data, to which a DL HAR model must be robust to be reliable in healthcare, lifestyle, and pervasive health moni-

toring applications. While it is known that data distribution affects the performance of DL models in general, the effects of the specific variabilities affecting IMU sensors in the performance of DL HAR models are unknown. Therefore, it is crucial to understand how distribution shifts in data due to variability affect the performance of DL HAR models and how to incorporate robustness into their performance evaluation.

1.1 Objectives

This study focuses on three major objectives to allow a deeper understanding of the effect of variability on DL HAR models.

1. Create a dataset designed specifically to study the effect of orientation, position, device, and subject variabilities in isolation from each other.
2. Quantify the effect of subject, orientation, position, and device variabilities using F1 score as a performance metric.
3. Measure the relationship between the impact of variability and the shift in data distribution by using Maximum Mean Discrepancy (MMD)[20].

Following this approach, this work aims to understand the data shift due to device and wearing variability and study their effects on DL HAR models.

1.2 Thesis Structure

This thesis is organized as follows: Chapter 2 (Related Work) provides a literature review that was performed for this study. Chapter 3 Introduces a new dataset named the HARVAR dataset that was collected for the facilitation of this study. Chapter 4 Explains in detail the methodology followed by this study to isolate and study different types of variabilities in HAR. Chapter 5 Depicts the results of this study, including a Gender-Based Analysis + and a case study using the REALDISP

dataset. Chapter 6 concludes the study and provides study implications, limitations, and summer and future work.

1.3 Publications

1. Preliminary results from this thesis have been published in [37].

2. Chapters 4, 5, and 6 are partially adapted from a publication [38] in Review at Sensors.

Chapter 2

Related Work

Human Activity Recognition (HAR) using wearable sensors involves classifying sensor signals into distinct human activities, with applications spanning healthcare and lifestyle monitoring. This chapter reviews existing DL HAR models, examining the factors contributing to their high reported accuracies and narrowing our focus to three DL HAR models selected for this study based on prior performance. Next, we discuss limitations in HAR evaluation methods, particularly those arising from data heterogeneity. To contextualize these challenges, we look at how data variability is addressed in other domains and review related research on the impact of real-world variabilities on HAR model performance.

2.1 Deep Learning Models in HAR

Traditional machine learning (ML) models classify activities by utilising features that are calculated from sensor data [13]. This requires significant effort in feature engineering, where selecting the right features is crucial for classifying activities [56]. However, this manual process was largely automated with the advent of deep learning (DL) techniques. DL models, particularly in Human Activity Recognition (HAR), can automatically handle feature extraction, making feature engineering unnecessary [51].

One of the most widely used DL architectures, Convolutional Neural Networks (CNNs), originally gained popularity in image recognition and computer vision [18, 73]. CNNs extract features by applying filters to the input data, identifying local dependencies between adjacent pixels, and using these extracted features for classification. Zeng et al. [85] demonstrated that this ability to capture local dependencies is also highly relevant for HAR, where dependencies between adjacent IMU readings can reveal important patterns in human movement. Furthermore, just as CNNs help achieve scale invariance in image recognition (where variations in color intensity do not affect recognition), they help in HAR by ensuring that the intensity of an action does not skew the classification. Early DL models in HAR used multi-layered CNNs to extract features from segments of IMU signals for activity classification [80].

On the other hand, recurrent Neural Networks (RNNs), another DL architecture, are sequential neural networks designed to learn temporal relationships in input data. Initially developed for language models [47], RNNs excel at capturing sequential dependencies within text, but they struggle with long sequences due to the vanishing gradient problem [27]. To overcome this, Long Short-Term Memory (LSTM) [70] and Gated Recurrent Unit (GRU) networks were introduced, which use gate mechanisms to manage long-term dependencies more effectively. LSTM has been used for HAR in pure LSTM models [75] where the researchers selected features manually and then trained the LSTM model to identify the activities based on the features. Unlike CNN, LSTM focuses on extracting temporal features and therefore requires researchers to still utilise manual feature engineering.

Combining CNNs and LSTMs addresses each architecture's limitations, creating a model that effectively captures both local and temporal features, which is particularly beneficial for sequential data. CNNs excel at extracting local features, while LSTMs capture temporal dependencies across sequences. This combination has shown significant improvements in various applications. For example, in gesture recognition, CNNs are used to extract features from individual video frames, while LSTMs capture the dependencies across frames, resulting in enhanced accuracy [55]. Similarly, this hybrid approach has been applied successfully to speech recognition, where CNN-LSTM

models show a 4-6% performance improvement over RNN models, which tend to retain less long-term information [81, 67, 17].

In the context of HAR, Ordóñez and Roggen [52] applied this combination of CNN and LSTM layers to classify activities from continuous IMU sensor data. Many of the recent DL HAR models fall into homogeneous and hybrid categories as explained in [58]. Homogeneous models, such as those in [29] and [23], rely solely on CNN or RNN architectures, while hybrid models combine CNNs with sequential networks, including RNNs [42], LSTMs [52], and GRUs [22]. The effectiveness of hybrid models stems from the complementary strengths of CNNs and sequential networks. CNNs excel at extracting local features and spatial patterns from input data, while sequential networks such as RNNs, LSTMs, and GRUs specialize in identifying temporal dependencies over time. By leveraging both local and temporal information, hybrid models can classify activities with greater accuracy, making them especially well-suited for HAR tasks, where both local features and long-term temporal relationships in sensor data are crucial. For instance, Ordóñez and Roggen [52] reported a 6% improvement in activity classification when using a CNN-LSTM hybrid model compared to architectures relying solely on CNNs. Given the superior performance of hybrid models, this study will focus exclusively on these models, which represent the state-of-the-art (SOTA) in HAR.

While multiple architectures have been proposed for HAR, we chose three hybrid DL-HAR models: DeepConvLSTM [52], TinyHAR [87], and Attend and Discriminate [1]. These models were selected due to their unique feature extraction and temporal information processing approaches, as detailed in Table 2.1. DeepConvLSTM [52] serves as a representative model for most hybrid DL HAR models, combining CNN with LSTM to extract local and temporal features. Originally, DeepConvLSTM used two LSTM layers, as proposed by Karpathy et al. [36], but **?**] demonstrated that a single LSTM layer performs better in most cases, which is why we will use a shallow DeepConvLSTM in this study. Both Attend and Discriminate [1] and TinyHAR [87] employ attention mechanisms [74] to improve feature extraction, with TinyHAR additionally optimizing the model to be lightweight.

DeepConvLSTM	Attend and Discriminate	TinyHAR
Feature Extraction: (Extraction of local or short time features)		
Four 1-dimensional convolution layers with a kernel size of 5, a stride of 2, and 64 filters are used to extract local features from input data.	Local feature extraction is done in two steps: First, the data is processed through 4 one dimensional convolution layers with a kernel size of 5, stride 2, and 64 filters. Second, a transformer encoder block comprising a self-attention and two fully connected feed-forward layers encode channel interaction. Where channels are the readings from various sensor modalities.	Local feature extraction is done in three steps: First, the data is processed through 4 one dimensional convolution layers with a kernel size of 5, stride 2, and 20 filters. Second, a transformer encoder block comprising a self-attention and two fully connected feed-forward layers encode channel interaction. Third, a fully connected layer fuses the cross-channel interaction information.
Temporal Information Extraction: (Extraction of features over the entire time window)		
A single LSTM layer with 128 cells is used to extract temporal features.	Temporal information is extracted in two steps: First, a single GRU layer with 128 cells extracts temporal features. Second, a self-attention layer is used to highlight important temporal features.	Temporal information is extracted in two steps: First, a single LSTM layer with 40 cells extracts temporal features. Second, a self-attention layer is used to highlight important temporal features.

Table 2.1: The three SOTA DL HAR models used for this study and their key architectural differences.

2.2 Limitations of Current HAR Evaluation and Data Heterogeneity

DL HAR models offer strong performance, but they have largely been tested on small, constrained datasets. DL models aren't fair, and their performance relies on the training data used [54]. It is often assumed that DL architectures, which perform well in image and language models trained on large datasets, will also perform effectively on the smaller datasets typically found in HAR [53], simply because they show good performance on the training and test sets used [52]. However,

in ML-based HAR models, the lack of diversity in data is recognized as a significant issue [24]. For example, research in slip-detection HAR highlights the limitations of small datasets with few participants, noting that current data prevents the development of a generalized model effective across diverse populations [78].

Real-world variability, such as human behavior or hardware differences, leads to distribution shifts within the data, adversely affecting model performance [77]. A distribution shift occurs when the distribution of features or class boundaries changes between the training and testing data, often due to real-world factors. This violates the common assumption that training and testing datasets follow the same distribution, leading to potential performance degradation in machine learning models when applied in real-world scenarios [49]. Although the impact of these shifts has been studied in domains like image recognition [71] and audio processing [34], showing negative effects, they have not been thoroughly explored in HAR.

In the audio domain, where the input signal is continuous and similar to IMU sensor data, Johnson and Grollmisch [34] studied the effect of distribution shift on DL models used to classify industrial sound and found performance drops to be related to the changes in the distribution of the data. Distribution shifts accounted for 9-10% drops in the DL model performances.

In the image domain, Taori et al. [71] assess the robustness of image classification models by evaluating models not based on their accuracy but by comparing the change in performance when testing the model with two test sets: One test had no distribution shift, and another test set had a distribution shift. Even with extensive training data, they found that DL models show a high susceptibility and corrupted classifications due to distribution shifts.

Subject variability has been shown to impact the accuracy of clinical and commercial actigraphy devices, as noted in Danzig et al. [16]. The study found that while these devices may perform reasonably well across larger populations, they are often unreliable for individual assessments due to factors like gender and age, which affect performance. Comparing actigraphy data to polysomnography, the study revealed discrepancies in recorded sleep timings, highlighting that actigraphy is not sufficiently accurate for precise, individual-level health assessments.

2.3 Measuring Distribution Shifts

Heterogeneity in data can be measured by assessing the shift in data distribution between the training and test sets, which helps quantify the impact of variability during model evaluation. Maximum Mean Discrepancy (MMD), as introduced by [20], is a kernel-based method that represents data into a Reproducing Kernel Hilbert Space (RKHS) to measure the differences between distributions in a high-dimensional, feature-transformed space. This allows MMD to detect non-linear differences between data distributions, making it effective for comparing multi-dimensional data like IMU sensor signals.

MMD has been used in various domains, such as image processing for unsupervised grouping in autoencoders [88] and domain adaptation by aligning image data distributions [5]. These applications underline MMD's ability to quantify distribution shifts, and in this research, MMD allows us to quantify shifts between the train and test sets of wearable sensor data. This helps us obtain a nuanced understanding of the change in DL model performance to the shift in data distribution.

Alternative methods, such as the Kullback-Leibler (KL) Divergence [41] and the Kolmogorov-Smirnov (KS) test [3], were less suitable for this research. The KS test is limited to single-dimensional data, whereas accelerometer data is three-dimensional. KL Divergence, on the other hand, relies on parametric assumptions that can be restrictive for wearable sensor data. In contrast, MMD is non-parametric and does not assume specific distribution forms (for example, Gaussian), making it better suited to capture the variability in real-world IMU sensor signals, even when noise or non-standard distribution shapes are present. This flexibility and the ability to process multi-dimensional data makes MMD a fitting choice for evaluating distribution shifts in DL HAR models.

2.4 Variability in HAR

Orientation, position, and device variability have been identified as a problem in HAR [61, 68], but they have either not been studied in isolation or not in the context of DL models. This has limited understanding of how these variabilities affect data distribution shifts and model perfor-

mance. We address this gap by analyzing the performance changes of DL-HAR models under isolated and combined variabilities and explaining the performance drop using MMD to measure data distribution shifts.

Orientation Variability in HAR has been studied in Yurtman and Barshan [82], where significant but varying drops in the performance of ML models were noted when the test data was subjected to random rotation. They found that some datasets experienced a 30% drop in accuracy due to rotation, while others showed no change. The paper did not explain this varying effect of orientation variability but focused on methods to reduce its impact.

Gil-Martín et al. [19] studied the effect of orientation change on Convolutional Neural Networks (CNNs). The baseline performance is obtained by training and testing the network with the original data from six public HAR datasets. 45° of orientation changes were induced via matrix transformation on the test set to measure the performance after the orientation change. They found that rotation transformation caused the model accuracy to drop by 2% to 11%, depending on which datase..0t was being used to train the ML models.

Orientation and Position variability for devices such as earbuds, which are fixed in position, can only induce orientation variability, as shown by Min et al. [48]. They found that orientation changes increased the Euclidean distance between the IMU data collected from different sessions for the same person and between different people and their habits. Positional variability due to sensor placement was studied on animals in Ahn et al. [2], where changing the sensor position from the back to the neck on dogs and horses led to a significant drop in the performance of unsupervised models for Animal Activity Recognition (AAR).

Najadat et al. [50] investigated the effect of device variability on HAR classification by training a Recurrent Neural Network model using smartphones and testing it with data from smartwatches. This resulted in an accuracy of 45%, nearly half the accuracy of other scenarios based on participant-wise train-test splits. This highlighted the impact of device variability, although it was mixed with the effect of position variability, as smartwatches were worn on the wrist while smartphones were worn on the waist.

Most research on the impact of variability in HAR has not isolated the effects of different variabilities. In some cases, the effects of multiple variabilities are combined, making it difficult to distinguish their contributions [50], while in others, variability is artificially induced [19]. In addition, the effects are not measured correctly as the study by Koh et al. [40] recommends measuring the performance drop by testing on the same distribution to isolate the effect of the distribution shift. Therefore, this research will use the latter approach to evaluate the effects of variability. A focused study that isolates orientation, position, device, and subject variability has not been conducted on DL HAR models. To achieve this, a new dataset is required to study each type of variability in isolation. In the next chapter, we discuss the HARVAR dataset, which was acquired to facilitate this study.

Chapter 3

HARVAR Dataset

To investigate how variability affects DL HAR models, we first needed annotated sensor data that includes ideal conditions and scenarios with variability in position, orientation, device type, and participant characteristics collected at the same time. However, as this chapter reveals, no dataset comprehensively covers these variability conditions, and therefore, we collected our own dataset to facilitate this study.

Wearable sensor-based HAR data is generated from IMU sensors like accelerometers, gyroscopes, and magnetometers. Researchers collect labeled IMU HAR data by having participants wear sensors and perform specific tasks, either video recorded or monitored in real time by a researcher who manually annotates the start and end of activities. After data collection, when multiple sensors are used, manual synchronization is often required due to slight variations in the internal clocks of each sensor. These discrepancies, especially in high-frequency data, can lead to misalignment. Once synchronized, the data is labeled using video recordings or the annotations made during the collection event. However, this labeling process often requires fine-tuning for precise alignment and can be labor-intensive and time-consuming.

Due to the effort involved in labeling, most labeled datasets are small and collected from local populations. Additionally, participants are often required to perform activities in a prescribed way, reducing natural variability in movements due to factors such as demographics, personal habits, or preferences. Furthermore, researchers typically attach IMU sensors in an ideal manner during data

collection, ensuring consistent placement across participants. In real-world applications, however, individuals may wear IMU sensors differently based on their comfort, habits, or style, leading to variability in sensor data that is not reflected in controlled lab environments.

These factors raise concerns about the real-world performance of HAR models, which are generally trained and tested on constrained datasets with limited diversity. The models may perform well in lab settings but struggle when applied to real-world scenarios with greater variability in how activities are performed or how sensors are worn.

To address this gap, the HARVAR dataset was created. HARVAR specifically focuses on capturing real-world variabilities, including differences in sensor placement, device type, and individual subject behavior. By collecting data without imposing restrictions on how participants perform activities or wear sensors, the HARVAR dataset isolates wearing, device, and subject variabilities, providing a more realistic and challenging benchmark for evaluating HAR models. This is crucial for improving the generalizability and robustness of DL models in real-world applications.

3.1 Existing Datasets and their limitations

Due to several limitations, many existing benchmark datasets in HAR, such as SKODA[84], OPPORTUNITY[14], WISDM[76], HHAR[9], Daphnet[62], PAMAP2[59], and DSADS[8]l, do not fully meet the requirements for studying the effects of real-world variabilities on DL models. For example, the SKODA dataset is collected from just one participant, severely limiting its ability to evaluate subject variability. A model tested on data from a single individual cannot effectively represent diverse populations, making it inadequate for real-world applications where participant variability is a key factor.

The OPPORTUNITY dataset [60], though comprehensive with data from four participants and 24 sensors per participant, lacks diversity in key aspects. Subject variability is limited due to the scripted activities protocol, which restricts natural movement variations and does not fully reflect real-world scenarios where individuals perform activities with more variation. While the dataset does include a diverse array of sensors—such as custom Bluetooth accelerometers, gyroscopes,

Sun SPOTs, InertiaCube3, a Ubisense localization system, and a magnetic field sensor—enabling analysis of position variability, it does not specify sensor orientations. This lack of orientation information makes separating position from orientation variability difficult. Additionally, the dataset does not place different sensor types in the same position, limiting the ability to study device variability in isolation from other factors.

The WISDM dataset [76] collects data from 51 participants, a larger group, but it has other limitations. It uses two devices—a smartphone (in the participant's pocket) and a smartwatch (on the wrist)—which creates a confounding factor. Any position variability is tied to device variability, and subject variability is intertwined with position and device changes. Additionally, the study does not clarify whether the participants were free to perform the activities naturally or were following scripted motions, making it difficult to assess the true variability in motion.

Some datasets, such as REALDISP [57], attempt to study real-world variability. In REALDISP, participants wear sensors themselves in one stage, and researchers place sensors in an ideal position in another stage. This design captures the difference between real-world and ideal settings. However, since these factors can be mixed, it still doesn't allow for the isolation of specific types of variability, like sensor orientation or subject variability. REALDISP also utilizes only one type of sensor, so this dataset can not be used to study device variability. Therefore, the dataset does not allow researchers to examine the impact of variability in isolation nor allow us to understand device variability.

The HHAR dataset [10] collects data from multiple smartphones positioned in a hip pouch and smartwatches worn on the dominant hand, enabling analysis of device heterogeneity. However, this dataset lacks coverage for other types of variability, such as sensor orientation and position. Additionally, it does not specify the orientation or sensor axis alignment for smartphones and smartwatches. This limitation makes it challenging to discern whether the observed device variability effects are confounded by orientation variability.

While existing datasets provide valuable insights into human activity recognition (HAR), they do not allow for the isolation and study of specific variabilities—such as subject, device, position, or orientation—needed to understand their isolated effects on model performance. Many of these

datasets rely on small, constrained populations following a scripted protocol, which limits subject variability. Moreover, they often lack clear distinctions among different types of variability or do not provide enough information to ensure variabilities are isolated, making it challenging to determine how each variability independently impacts HAR models. Consequently, these datasets are insufficient for evaluating how models are affected by each variability.

To address these limitations, we developed the HARVAR dataset, specifically designed to isolate and analyze multiple types of variability simultaneously under the same context. Our study protocol enables data collection that captures subject, device, position, and orientation variability, with data collected from eight sensors across 16 participants, which allowed them to perform activities freely without any constraining protocol. This dataset allows for assessing the effects of individual variabilities on HAR model robustness and generalization.

3.2 Sensors and Sensor Placement

For data collection, we used Empatica Embrace Plus watches and Bluesense sensors [63], both of which feature an accelerometer, gyroscope, and magnetometer. These sensors were chosen because they provide direct access to raw data, unlike many other smartwatches offering only processed outputs, such as step counts or general activity levels, rather than raw IMU data. Table 3.1 outlines these sensors' specific characteristics and placements, highlighting their different sampling rates, which adds another layer of variability to our analysis.

Two Empatica Embrace Plus watches were positioned on each participant's right and left wrists in the recommended placement, just above the wrist bone, with a spacing of 1-2 finger widths. Multiple Bluesense sensors were used: one on the left wrist in an ideal position, two on the right wrist—one placed ideally, and another rotated 45 degrees along the z-axis. Further Bluesense sensors were attached to the participants' upper right arm, upper left arm, and torso to gather comprehensive motion data from various positions. The placement of the sensors is further illustrated in the Figure 3.1

The sensors were attached strategically to enable a comparative analysis highlighting the impact of different variabilities. The reasoning for their placement is as follows:

- Sensors bluesense-RWR1 and bluesense RWR2 preserve all aspects, such as position and device, but vary in orientation by 45 degrees on the z-axis. This reveals the change in the data collected from a slight orientation change.
- Sensors empatica-right and empatica-left are placed in the ideal positions on the left and right wrist, preserving device type and orientation but not position. This pair allows us to understand the difference in data when people change the wrist on which they wear the IMU sensor.
- Device pairs such as bluesense-LWR and bluesense-LUA or bluesense-RWR1 and bluesense-RUA preserve device type and orientation but not position.
- Each wrist has two devices in the same position and orientation. For example, bluesense-LWR and empatica-left on the left wrist preserve position and orientation but not device type. This pair highlights the difference in data collected from two devices of different types under an unchanged context.
- Finally, the torso sensor blue sense-TRS serves as a good reference IMU sensor that does not experience any drastic movement due to the fast movement involving hands in human activities.

3.3 Activity Protocol for Data Collection

Once the sensors were properly attached, the participants were asked to perform 2 activity routines. First, they were asked to walk on a treadmill at varying speeds, and then they were required to prepare a simple salad and eat it. Walking on a treadmill is a controlled periodic activity, as we can control the speed at which the participant walks, and the walking motion does not show large variability between people. Salad preparation, on the other hand, is a complex activity with many

Sensor Type	Sensor Position	Sensor Name	Sensor Code	Sampling Frequency
Bluesense	Right Wrist (No rotation)	bluesense-RWR1	BR1	100Hz
Bluesense	Right Wrist (45 rotation)	bluesense-RWR2	BR2	100Hz
Bluesense	Left Wrist	bluesense-LWR	BL	100Hz
Bluesense	Right Upper Arm	bluesense-RUA	BRU	100Hz
Bluesense	Left Upper Arm	bluesense-LUA	BLU	100Hz
Bluesense	Torso	bluesense-TRS	BA	100Hz
Empatica	Right Wrist	empatica-right	ER	64Hz
Empatica	Left Wrist	empatica-left	EL	64Hz

Table 3.1: T nformation on their sampling rate and placement.

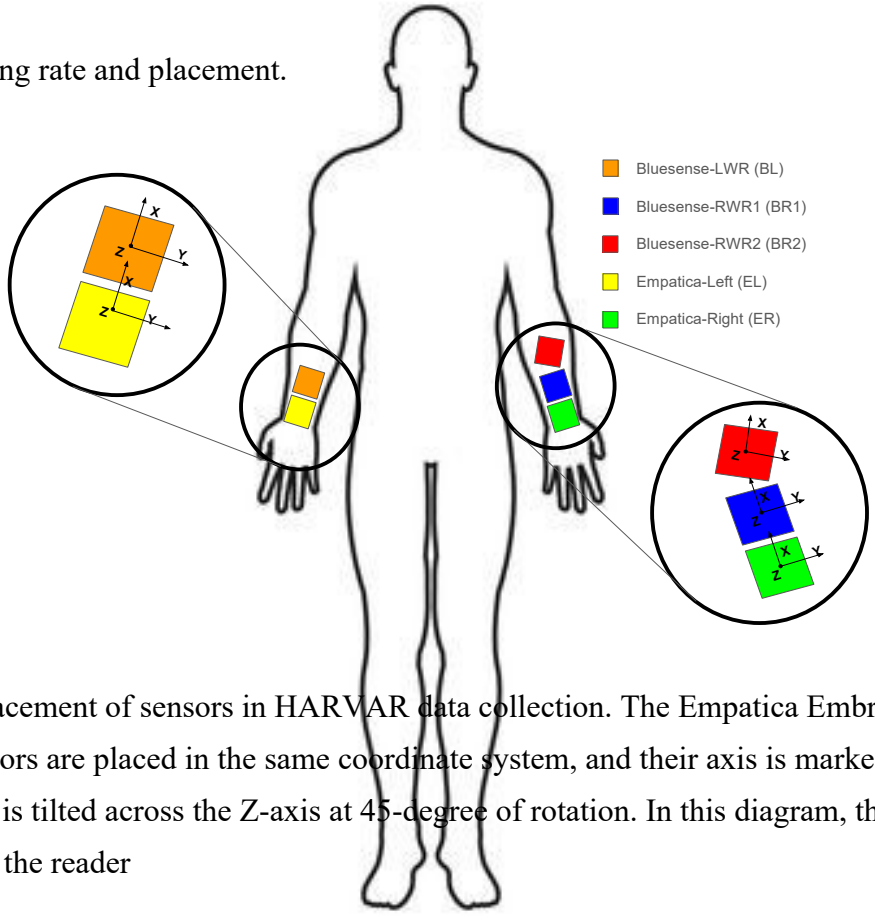


Figure 3.1: Placement of sensors in HARVAR data collection. The Empatica Embrace Plus and Bluesense sensors are placed in the same coordinate system, and their axis is marked. BR2, marked as red, is tilted across the Z-axis at 45-degree of rotation. In this diagram, the person is facing towards the reader

intricate movements and was selected to highlight the differences in performing activities due to habits and demographic changes.

First, participants walked on a treadmill at five speeds: 3.2 km/h, 4 km/h, 4.8 km/h, 5.6 km/h, and 6.4 km/h. Participants were instructed to walk naturally and comfortably without strict guidelines to ensure they didn't constrain themselves in moving a certain way. Each speed lasted approximately 2 minutes, totaling 10 minutes of walking, accounting for transitions between speeds.

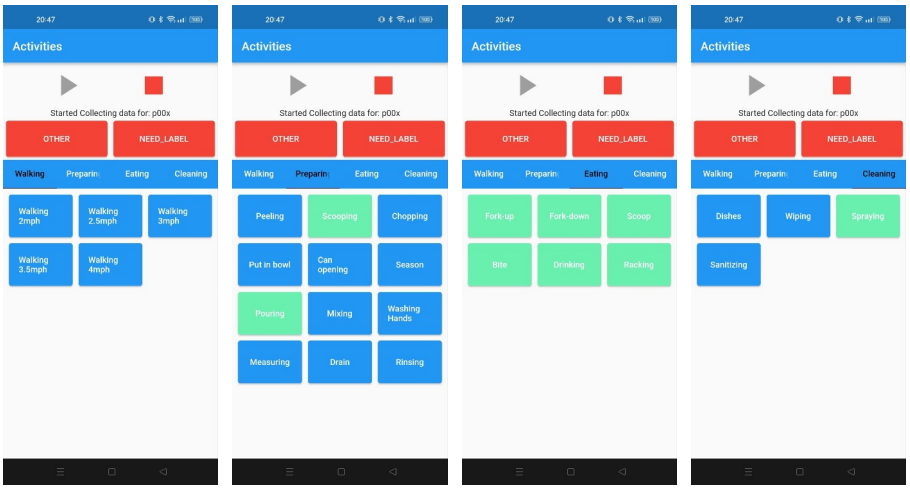


Figure 3.2: Screenshots of the data annotation application used to roughly annotate the activities. The application was used by researchers who observed the activities done by the participants.

Following the treadmill walk, participants were asked to perform a simple salad preparation task. While the participants were given a recipe to follow, they were allowed to perform the steps in whatever way they found comfortable. The salad preparation involved washing vegetables, seasoning, mixing ingredients, eating the salad, and washing dishes afterward. This task, lasting approximately 20 minutes, was intended to capture more complex, naturalistic motions, providing variability in both the method and duration of the actions.

During data collection, the walking and salad preparation activities were video recorded to enable manual labeling of the activities after data collection. This video recording was needed to map IMU sensor data accurately to specific activities. A researcher was assigned to manually label the activities in real-time using a custom-developed mobile application to assist with the labeling process. This application, built in-house, was designed to create time-stamped JSON labels for each activity performed. Figure 3.2 shows a screenshot of the application.

While the participants were performing their activities, the researcher would observe them and use the app's interface to quickly mark the start and end of specific activities. This annotation phase was rough and served only as an initial guide for more detailed labeling during the post-collection data processing, which is discussed further in this work.

3.4 Sensor Data Transmission, Storage, Syncing, and Labeling

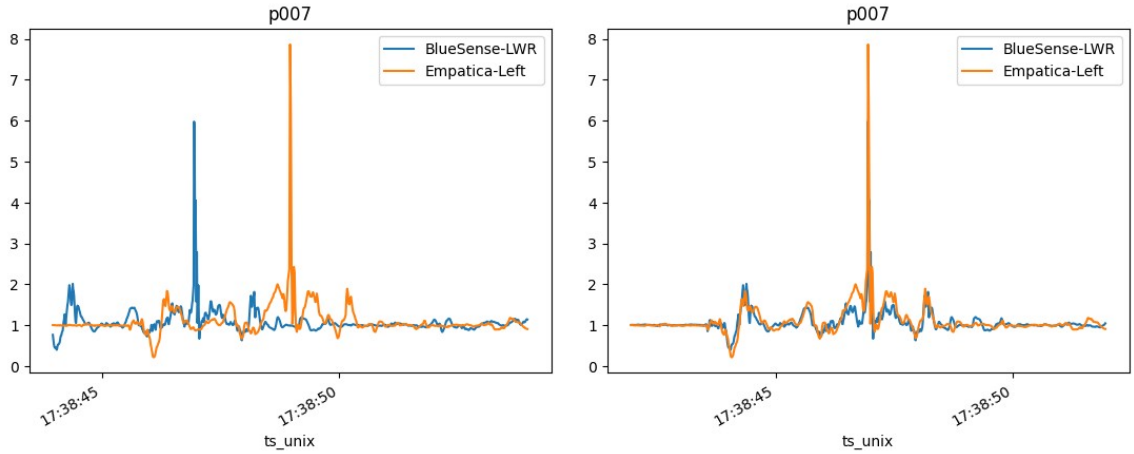
The sensor data for the Bluesense and Empatica devices were collected via Bluetooth during the experiments, but these sensors operated differently. The Bluesense sensors are connected directly to a PC, and the data is saved in real time during the experiments. In contrast, the Empatica Embrace Plus sensors transmitted data to a mobile application, which then processed and uploaded the data to the Empatica cloud storage for future access.

Once data collection was complete, the sensor data quality from all 16 participants was verified, and no missing data was identified. However, synchronizing the data between sensors was essential because data from multiple sensors was collected simultaneously. Although each sensor recorded data with time stamps, the internal clocks on each device were not perfectly aligned, resulting in slight shifts in time stamps—ranging from a few milliseconds to several seconds—across different devices.

To address this misalignment, we implemented a unique marker in the accelerometer data: a strong, deliberate clap. Participants were asked to perform an exaggerated clap once before the walking activity and once before the salad preparation. The accelerometer data captured from these claps produced distinct peaks, which were easily identifiable in the data. The magnitude peak between the Bluesense-LWR and the Empatica-Left sensors for Participant 7 are shown in Figure 3.3. We synchronized the data from all devices by aligning the peak magnitudes of these claps across all sensors, using the Empatica Left sensor as the reference point.

The activity labeling process was conducted using the open-source Label Studio [72]. The annotation involved identifying each activity's start and end points within the dataset. Initially, the video recordings were synchronized with the IMU sensor data using the clap artifact previously used for sensor alignment.

The annotations provided by the mobile application were largely accurate for the walking activities, given the continuous and long nature of the walking segments. As such, these initial labels required minimal adjustment to ensure consistency with the IMU data.



(a) Before data synchronization **(b)** After data synchronization

Figure 3.3: The figure on the left depicts the Bluesense-LWR and the Empatica-Left sensor magnitude before their timestamps were synchronized. On the right is the depiction of signals after synchronization has been performed. This data is of Participant 7 and shows the clap that was performed before the walking activity.

In contrast, the salad preparation activity needed manual labeling due to the short and varying lengths of actions. Video recordings were reviewed to mark the start and end of activities, such as chopping, washing dishes, eating, and washing hands. While the mobile application provided rough time stamps, these were adjusted based on the video to ensure that the activity boundaries were precisely marked.

After completing the annotation process, the data was divided into two categories: walking and cooking (salad preparation). The IMU data was further separated based on whether it came from the Bluesense or Empatica sensors for each sensor and each participant. The file structure in which the data was stored is depicted in Figure 3.4

3.5 Data Overview

The dataset includes 16 participants from diverse age groups, as shown in Table 3.2. The mean age is 42, with a standard deviation of 20 years. There were 5 participants aged 60 and above, 3 participants aged 30 to 60, and 8 participants between 20 to 30. The data consists of 9 male participants,

Figure 3.4: The file structure in which the HARVAR Data is stored. In this figure, pxxx can be any participant number, such as p001 to p016.

with a mean weight of 74 kg and a standard deviation of 13 kg, and 7 female participants, with a mean weight of 62 kg and a standard deviation of 13 kg.

The duration of all activities performed during the data collection event is provided in seconds in Table 3.3. The total duration of walking (at speeds 4.8 km/h, 5.6 km/h, and 6.4 km/h) for all participants is slightly lower than 32 minutes ($2 minutes \times 16 = 1920 seconds$) as some participants did not complete the full 2-minute walk at higher speeds due to age or physical limitations. The combined treadmill walk for 16 participants was 8763 seconds long.

The salad preparation was more than double the length of the walking activity. In total, participants spent roughly 21614 seconds during the cooking activity. Therefore, on average, the salad preparation took around 22 minutes per participant. So far, the activities of Biting, Washing Hands,

Table 3.2: Inforamtion about the 16 participants of HARVAR Dataset.

ID	Age	Sex	Weight (Kgs)	Holding Sidebar
1	59	m	83.9	no
2	74	f	65.7	yes
3	60	f	49.8	no
4	71	m	79.0	yes
5	61	f	55.3	no
6	71	m	64.8	no
7	26	f	73.0	no
8	25	m	72.5	no
9	26	m	61.0	yes
10	47	m	89.8	no
11	23	f	53.0	no
12	21	m	55.0	no
13	24	f	74.8	no
14	35	f	86.2	no
15	29	m	73.0	no
16	26	m	95.0	no

and Drinking are the only labeled activities. Therefore in Salad preparation, only 18 minutes of data out of the total 360 minutes are labelled. The rest of the data is currently marked as 'Other' in Table 3.3.

Table 3.3: Duration of each labeled activity in HARVAR Dataset

Activity Name	Duration in Seconds
Walking on Treadmill	
Walking 3.2 kmph	1998
Walking 4 kmph	1981
Walking 4.8 kmph	1767
Walking 5.6 kmph	1630
Walking 6.4 kmph	1387
Salad Preparation	
Washing Hands	406
Biting	531
Drinking	157
Other	20520

3.6 Potential Use Cases of the dataset

The HARVAR dataset is designed to allow researchers to isolate and study various real-world variabilities such as orientation, position, device, and subject variability in IMU-based HAR data. Beyond that, the dataset is valuable for evaluating the robustness of HAR models, ensuring they are better prepared for deployment in real-world scenarios.

This dataset also provides insight into motion variability across individuals, particularly during complex tasks like salad preparation, where different people may perform the same activity in varied ways. Additionally, short-duration actions (gestures), such as taking a bite, create opportunities for studying activity recognition in applications related to monitoring eating habits. HARVAR can also be used for standard model training and evaluation in HAR. A diverse group of 16 participants and a range of activities allow for the assessment of model performance not only under variability but also in controlled scenarios, offering a comprehensive evaluation of model capabilities.

The HARVAR dataset includes activities, such as walking and drinking water, found in existing datasets like OPPORTUNITY and WSDM. This overlap allows HARVAR to be valuable in studying cross-dataset generalizability, where we can assess if a model trained on features from one dataset can accurately classify the same activities in another dataset, even when different sensor types are used. This type of evaluation is essential for understanding a model's robustness and adaptability across diverse data sources.

In this study, we use the HARVAR dataset to examine the effects of these variabilities and quantify the resulting shifts in data distribution. In the next chapter, we discuss how the HARVAR dataset is utilized to isolate orientation, position, device, and subject variability while preserving the context of the same experiment.

Chapter 4

Method to evaluate the Effect of Variability on Deep Learning HAR Models

The HARVAR dataset enables us to isolate key variabilities—device, orientation, position, and subject— providing a structured approach to assess their impacts on DL HAR model performance. This chapter explains the methodology for isolating the effects of each variability in DL HAR model performance using the HARVAR dataset. We then explain the method to quantify data distribution shift and draw a relationship between the change in DL model performance and the changes in data distribution. Finally, we explain how to study variability and data distribution shifts when multiple variabilities are at play to simulate real-world scenarios.

4.1 Experimental Protocol to Evaluate Model Performance under Variability

To understand the effect of each type of variability to be studied, we selected sensor pairs representing Device, Position, and Orientation Variability from the HARVAR dataset. We evaluated each model for selected pairs of sensors: once for the baseline scenario with no variability (in which the train and test data are of the same sensor, sensor 1) and once for the variability scenario (when the train data is of sensor 2 and the test data is of sensor 1). The pairs for position and

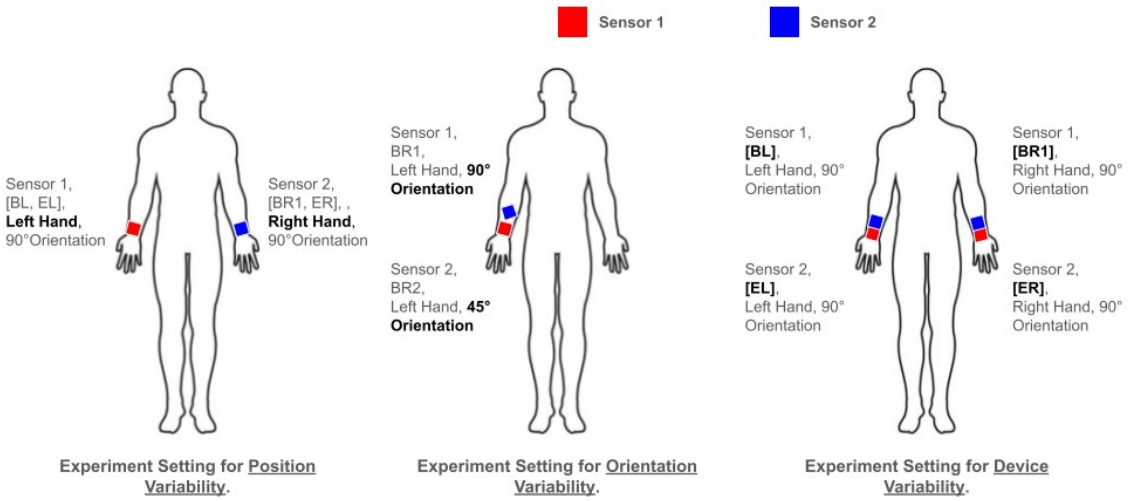


Figure 4.1: The experiment setting using the HARVAR dataset to evaluate the effect of device, position, and orientation variability. Where Sensor 1 and Sensor 2 are used in combination as a train-test pair to highlight variability. In these diagrams, the person is facing towards the reader.

orientation variability use the same type of device on different wrists (position) or the same wrist but with rotation (orientation), respectively. In contrast, we use the same position and orientation for device variability but different device types. These pairs are depicted in Table 4.1 from rows 1 to 8 and Figure 4.1. Notice that in each experiment, the test sensor is the same to isolate the effects of the variability following recommendations in [40]. Testing with a different sensor would combine the effects of the different testing distributions and the variability. We evaluate the effect of variability as the performance disparity, measured as the difference in F1-Score, between the two settings.

In our evaluation, we employed cross-validation using a leave-one-subject-out (LOSO) approach for each experimental setting. For instance, as shown in Table 4.1, in Experiment 1's variability setting, we trained each model on data from the Empatica-right sensor of Participants 2-16, testing on the Empatica-left sensor of Participant 1 for the first fold. By examining each LOSO fold—where the test data came from a unique participant—we observed the effect of subject variability on model performance. Specifically, we assessed the performance in the baseline

condition (where the training and testing sensors were identical, with no variability introduced) across each fold, which allowed us to highlight the changes in model performance due to subject-specific differences.

Table 4.1: List of Experiments conducted using the HARVAR Dataset.

Exp. ID	Variability type	Train Sensor	Test Sensor	Setting
1.	Position	empatica-right	empatica-left	variability
		empatica-left	empatica-left	baseline
2.	Position	empatica-left	empatica-right	variability
		empatica-right	empatica-right	baseline
3.	Position	BRW1	BLW	variability
		BLW	BLW	baseline
4.	Position	BLW	BRW1	variability
		BRW1	BRW1	baseline
5.	Device	BLW	empatica-left	variability
		empatica-left	empatica-left	baseline
6.	Device	empatica-left	BLW	variability
		BLW	BLW	baseline
7.	Orientation	BRW1	BRW2	variability
		BRW2	BRW2	baseline
8.	Orientation	BRW2	BRW1	variability
		BRW1	BRW1	baseline

4.1.1 Model Training

As mentioned, we evaluated three DL models: DeepConvLSTM, Attend&Discriminate, and Tiny-HAR, as depicted in Table 2.1

We trained the models on a simple binary classification task: identifying whether or not the subject was walking. Walking was chosen because it is simple and, as shown by Xochicale et al. [79], complex motions may differ significantly between individuals. By focusing on this controlled walking activity, we aim to minimize the impact of motion variability, which is further mitigated through LOSO cross-validation during testing.

The models were trained for one sensor at a time, with the input being the 3-dimensional accelerometer data. The training data was normalized in isolation from the test data using stan-

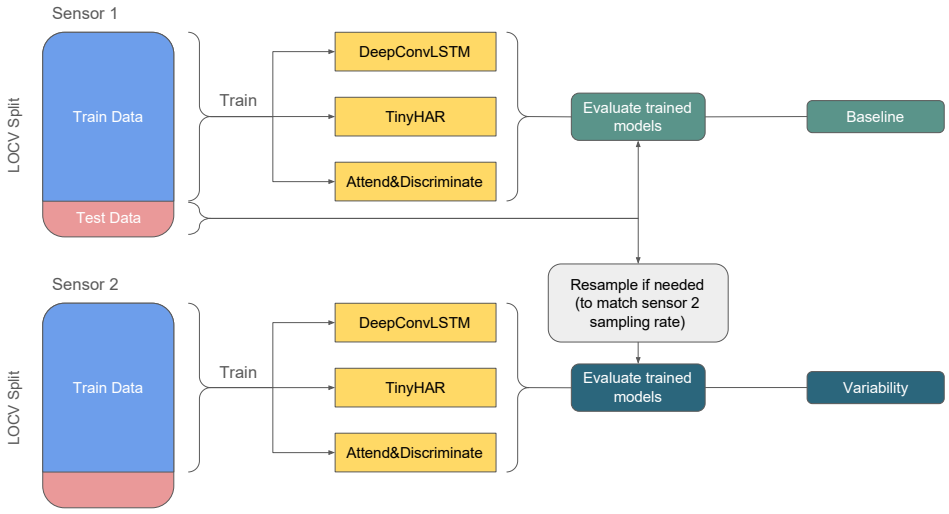


Figure 4.2: The process of evaluating the effect of variability using the HARVAR dataset.

dardization and then split into sliding windows. Datasets like Opportunity, Skoda, WISDM, and PAMAP2 advise window sizes to vary between 1 to 5 seconds [25, 45, 64], as they all compose of various activities, some short while others long and periodic. The research done by Janidarmian et al. [31] suggests using window sizes longer than 3 seconds for cases where only one sensor is being used (like in our case) but recommends reducing the window size if the activity is periodic (such as walking). Since we are concerned with the walking activity in the HARVAR dataset, we used a 2-second sliding window.

The sliding windows were shuffled and split into a 9:1 ratio of train and validation. No other form of pre-processing was used on the training or validation data to maintain originality. A weighted data loader was utilized during the training process as the samples of the ”not-walking” class outweighed the ”walking” class by a 2:1 ratio. Models were trained using a batch size of 256 over 150 epochs, with early stopping called after 15 epochs of no improvement over the validation set. The initial learning rate [66] was set to 0.001, using learning rate annealing [30] with patience of 7 epochs and a reduction factor of 0.1. The Adam optimizer [39] was utilized, optimizing based on the CrossEntropy [65] criterion. These hyperparameters were picked based on the hyperparameters used by [87] and [1].

Table 4.2: The computational complexity (in MACs) variance between the different models. The computational complexity depends on the model architecture and the sensor being used to train due to the difference in sensor sampling rates.

Model	Sensor	Computational Complexity (MACs)	Parameters
TinyHAR	Bluesense	$2.53 * 10^8$	24864
DeepConvLSTM	Bluesense	$2.19 * 10^7$	227138
Attend&Discriminate	Bluesense	$8.36 * 10^7$	297412
TinyHAR	EmpaticaEmbrace+	$1.53 * 10^8$	24864
DeepConvLSTM	EmpaticaEmbrace+	$1.33 * 10^7$	227138
Attend&Discriminate	EmpaticaEmbrace+	$5.12 * 10^7$	297412

Note that the model architecture remains consistent between all experiments. Still, since the Empatica and Bluesense sensors run at different sampling rates (as shown in Table 3.1), the model complexity varies depending on which sensor is used to train the model. This difference in model complexity is depicted in Table 4.2, and the model complexity is calculated as Multiple Accumulate Operations (MACs). A MAC involves performing a multiplication followed by an addition. It is commonly used to quantify the computational load in a DL model. Each operation (such as a convolution or matrix multiplication) can be broken down into a series of MACs, and the total number of MACs can be used to estimate the computational complexity of the model. If a model performs more MACs, it indicates that it is performing more computational work, suggesting that it is more complex and requires more computational resources. Conversely, a model that performs fewer MACs is less complex and more efficient in computation.

Bluesense sensors have a sampling frequency of 100 Hz, and Empatica sensors have a sampling frequency of 64 Hz, as shown in Table 3.1. Since we do not employ resampling before the training process, the input window size for models trained using Empatica sensors is 128, while the input window size for models trained using Bluesense sensors is 200. As detailed in the following subsection and shown in Figure 4.2, resampling will only be performed using interpolation during the testing phase when the train and test sensors have different sampling rates. Since every layer in the model depends on the input size, the models trained using Bluesense sensors are more complex than those trained with Empatica.

4.1.2 Model Performance Evaluation

Once the models are trained, they are tested using the data from the participants left out during the LOSO training process. Multiple tests use data from different sensors each time, achieving the train-test pairs depicted in Table 4.1. The F1 score is used to compare performance across various experimental settings.

For device variability, due to the different sampling frequencies of the devices, the test data is interpolated to match the sampling frequency of the training data before being input into the model. When a model is trained with data from a particular sensor, the input data is not resampled, and the original data is used directly. For example, if the model is trained using the Bluesense sensor, which has a sampling rate of 100Hz, it would expect the input for a 2-second window to consist of 200 samples. However, if we want to test the model with Empatica sensor data, which has a sampling rate of 64Hz and produces 128 samples for the same 2-second window, the Empatica data must be resampled to match the Bluesense data's length of 200 samples before being fed into the model.

4.1.3 Measuring Variability with MMD

Maximum Mean Discrepancy (MMD) is a kernel-based statistical test used to determine the similarity between data distributions. We hypothesized that variability introduces a distribution shift in the data, contributing to the observed effects on the performance of DL HAR models. Our study employed the multiscale kernel for MMD with a bandwidth range of [0.2, 0.5, 0.9, 1.3, 1.5, 1.6]. A higher MMD value indicates a greater difference or shift in data distribution, whereas a smaller MMD value suggests a smaller shift. This approach helps quantify the impact of variability on the data distribution and, consequently, on model performance.

We calculate the MMD value between the train and test splits used to evaluate the DL HAR models, both with and without variability. To maintain consistency in MMD calculation, we perform the same data preprocessing steps used in the DL HAR Model Training 4.1.1 and Performance Evaluation 4.1.2. The MMD is computed only for labeled data, excluding null or negative

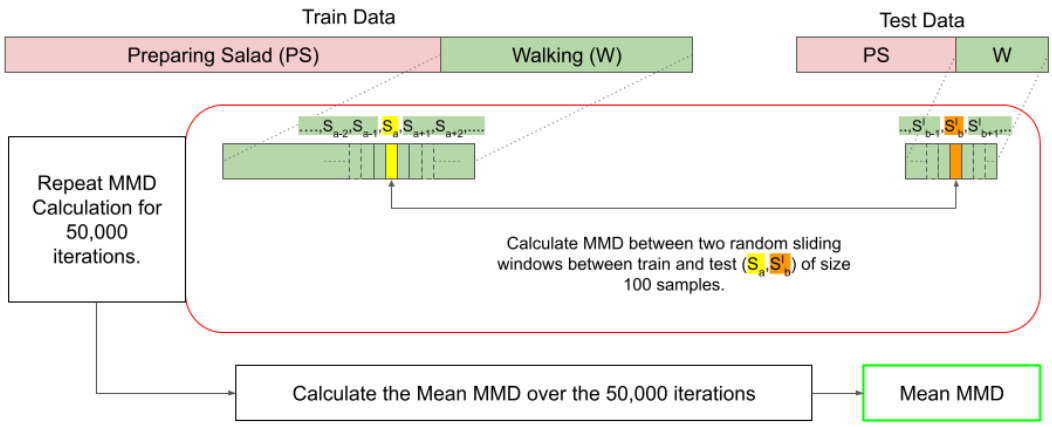


Figure 4.3: The process of calculating the MMD between the train and test data of the HARVAR dataset.

classes. For the HARVAR dataset, MMD is calculated exclusively for the walking class, omitting "not walking" activities due to their heterogeneous nature.

Typically, MMD is calculated as a one-shot process between two data distributions, but this requires matrix multiplication and becomes computationally unfeasible for long data sequences. In our case, the walking activity spans roughly 10 minutes per participant. Bluesense sensors collect data at 100 Hz and Empatica Embrace Plus at 64 Hz, resulting in approximately 60,000 and 38,400 data points, respectively. Performing MMD calculations on datasets of this size in one shot becomes impractical.

We adopt an iterative approach to address this, as illustrated in Figure 4.3. We randomly select a 100-sample window from the training and testing datasets, compute the MMD between them, and repeat this process over multiple iterations. Repeating this for 50,000 iterations, we calculate the average MMD, which we use as the final MMD value representing the MMD between the training and testing datasets. Since the MMD is calculated as a mean over 50,000 randomly selected pairs of windows of size 100 samples, it is a statistically valid quantification.

An iterative approach is suitable for continuous and repetitive actions like walking, as no unique artifacts in the walking data are critical for classification. However, it may not be appropriate for short-duration activities such as taking a bite, sipping water, or opening a door, where the complete time window of the activity is essential for accurate MMD calculation. In these cases, capturing

the entire activity sequence is necessary to account for specific, brief movements that are key to recognizing these actions.

4.2 Measuring Compounding Effects of Variability

While the HARVAR dataset allows us to measure the effects of each type of variability in isolation, real-life scenarios often involve a combination of these variabilities. We utilized the REALDISP [6, 7] dataset to study this compounding effect. The REALDISP dataset highlights the combined effect of wearing variabilities by comparing data collected from wearable IMU sensors placed ideally by researchers (Ideal) and data from sensors worn unsupervised by participants (Self). It was collected from 16 participants over two iterations for each participant. In the first attempt (Self), the participants wore the sensors without the guidance of the researchers to mimic the real-life placement of consumers who wear smart devices with IMU sensors. The second time in the 'Ideal' scenario, the IMU sensors were attached to the participants by the researchers in an ideal position and orientation.

In the 'Self' setting, the sensors' position and orientation differ from the 'Ideal' setting. The orientation can vary by as much as 180 degrees if worn upside down, as the sensors lack a reference for the "correct" orientation. Unlike the HARVAR dataset, the position variability here is more subtle, as it does not involve switching the sensor from one wrist to another. Instead, the variability comes from minor changes along the arm's length. For instance, depending on the participant's comfort, a sensor could be worn on the wrist or the forearm.

We conducted the experiments summarized in Table 4.3 to investigate how wearing variability, induced by the combined effects of position and orientation variability, impacts the performance of DL models. The selected scenarios represent various training and testing conditions commonly encountered in HAR model evaluation.

The first two scenarios compare the ideal case with a variability case. In the first scenario, both training and testing are conducted using 'Ideal' data, while in the second scenario, 'Self' data is used for training and 'Ideal' data for testing. These scenarios are analogous to experiments

Table 4.3: Experiments to evaluate compound effects of variability using the REALDISP Dataset.

Exp. ID	Scenario	Train Data	Test Data	Sensor
1.	Lab scenario, trained and tested with ideal data.	Ideal	Ideal	RLA/LLA
2.	Trained with variable data and tested with ideal data	Self	Ideal	RLA/LLA
3.	Trained and tested with variable data.	Self	Self	RLA/LLA
4.	Lab trained and tested on variable data.	Ideal	Self	RLA/LLA

conducted using the HARVAR dataset, featuring a non-variability scenario (Ideal vs. Ideal) and a variability scenario (Self vs. Ideal). By comparing the performance drop between these two scenarios, we aim to understand the effect of compounded wearing variabilities, such as orientation and position. It is important to note that this evaluation differs from HARVAR in that HARVAR features controlled variability, where the variability is consistent across all participants. In contrast, the variability in the REALDISP dataset varies from participant to participant, as the 'Self' data depends on how each participant wears the sensor.

The next two scenarios simulate real-world conditions faced when training DL HAR models. The third scenario reflects training and testing data collected in unconstrained, real-world conditions, allowing variability in both training and testing. The fourth scenario represents a situation where a model is trained on lab-collected ideal data and then deployed in real-world settings where user-induced variabilities can affect performance (Ideal vs. Self).

Each of the four scenarios is run twice, using data from the right lower arm (RLA) and once from the left lower arm (LLA). This ensures that our results are not biased by any differences in data caused by the dominant and non-dominant arms.

MMD is calculated between the train and test sets, similar to the approach used with the HARVAR dataset. However, unlike HARVAR, where we only utilized one walking activity, the REALDISP dataset includes 33 different activities for classification. To calculate the MMD in this case, we compute the MMD between the train and test sets for each activity separately, following the same process as HARVAR (Figure 4.3). We then average the MMD values for each activity, quantifying the distribution shift between the train and test.

Chapter 5

Results of Variability Evaluation

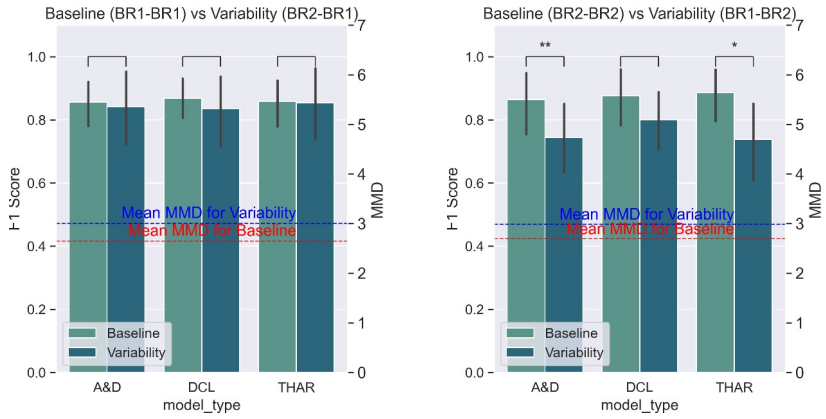
In this chapter, we present the results of our study. The chapter is organized into sections, each exploring a research question. We begin by studying the impacts of each data variability on model performance using the HARVAR dataset in section 5.1. We then use the Maximum Mean Discrepancy Metric (MMD) to explain performance differences across variabilities and participants in section 5.2. Using HARVAR data and the performance results of the three DL models, we perform a (Gender-Based Analysis) GBA+ analysis in section 5.3. Finally, in Section 5.4, we study the combined effects of variability using the REALDISP Dataset as a more realistic scenario.

5.1 Variability Impacts on Model Performance

We evaluated the impact of data variability on model performance by comparing the F1-Score difference on the baseline and variability settings, as explained in Chapter 4.1. Figures 5.1, 5.2 and 5.3 show the average and standard deviation of the F1-Score across all validation folds of each of the three evaluated models under no variability (dark green) and orientation, position, and device variability (light green) settings, respectively. The mean F1 score is calculated over the F1 score acquired from all participants during LOSO cross-validation. MMD values, also shown in these figures, will be explained in Section 5.2.

The significance of the difference in performance between the baseline and variability settings is tested using a T-test[12]. The null hypothesis posits no difference between the baseline and variability scenarios. This hypothesis holds if the p-value of the T-test is greater than 0.05. Conversely, if there is a significant difference in performance, the p-value will be less than 0.05, indicated by *, less than 0.01 by **, and less than 0.001 by ***.

5.1.1 Orientation Variability



(a) Orientation variability between BR2 and BR1 when the test sensor is BR1

(b) Orientation variability between BR2 and BR1 when the test sensor is BR2

Figure 5.1: Performance changes due to Orientation Variability. We show the average F1 score and average MMD values for each DL HAR model in the two experiments. Light green bars represent the no variability setting of each experiment, and dark green bars represent the variability setting. Asterisks represent the p-value of a paired t-test (*: $p - value < 0.05$, **: $p - value < 0.01$, ***: $p - value < 0.001$). Only two models in one experiment showed significant performance changes, but the F1-Score remains above 0.7.

Orientation variability due to the rotation of an accelerometer along one of its axes was tested using the BR1 and BR2 sensors, as shown in Figure 4.1. These sensors have a 45-degree rotation difference but are both on the right wrist.

Figure 5.1 depicts the model performance for experiments 7 (Figure 5.1b) and 8 (Figure 5.1a) in Table 4.1. When BR2 is the test sensor (Figure 5.1b), we do not observe any significant model performance changes due to orientation variability ($p > 0.05$ in a paired t-test). For any of the evaluated models, in contrast when BR1 is the test sensor(Figure 5.1a), we see a significant drop in the performance of the Attend&Discriminate model ($p < 0.001$) and in the performance of the TinyHAR model ($p < 0.05$). We do not see a significant drop in the performance of the DeepConvLSTM model ($p > 0.05$).

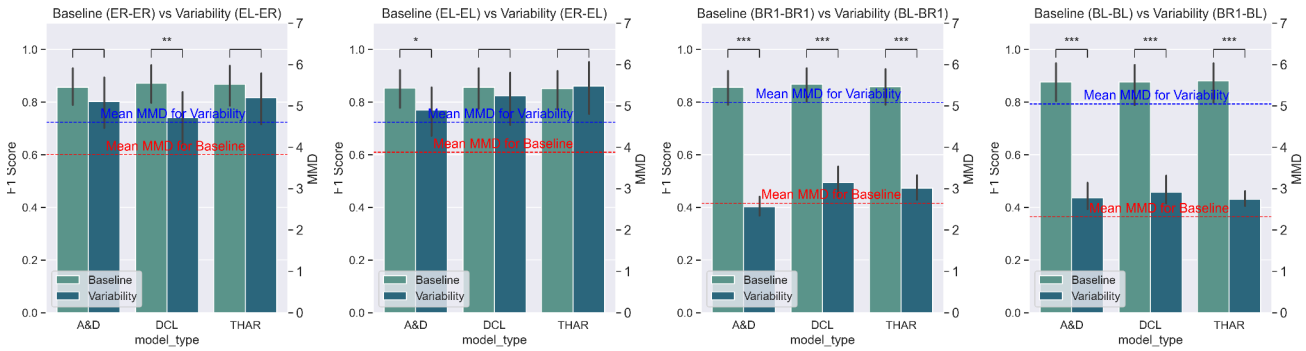
In both orientation variability experiments shown in Figure 5.1, the performance of the baseline setting remains similar (F1 score 0.86) for all models regardless of the test sensor. However, in the variability scenario, we see a difference in performance between the two experiments:

1. In Figure 5.1b, when the model is trained with the sensor BR2 (which is rotated 45 degrees) and tested with sensor BR1 (with no rotation), we see no significant drop in performance. In this variability experiment, the F1 score remains above 0.81 for all three DL models.

2. In Figure 5.1a when the model is trained with the sensor BR1 (which has no rotation) and tested with sensor BR2 (with 45 degrees of rotation), we see a significant drop in performance for two DL HAR models (Attend&Discriminate and TinyHAR). In this variability experiment, the F1 score is less than 0.8 for DeepConvLSTM and less than 0.75 for Attend&Discriminate and TinyHAR.

5.1.2 Positional Variability

Positional variability was evaluated across four experiments, as shown in Figure 5.2, each using pairs of the same device on different wrists. The results varied depending on the sensors being used. Comparing Figures 5.2a and 5.2b (Empatica), with Figures 5.2c and 5.2d (Bluesense), we observe a greater performance drop due to positional variability when using Bluesense sensors (mean F1-Score difference of 0.45 and $p < 0.001$) than when using Empatica sensors (mean F1-Score difference of 0.12 and p-value close to 0.05). Since the type of variability is the same and the DL model architectures are unchanged, the larger drop in performance can be attributed to the



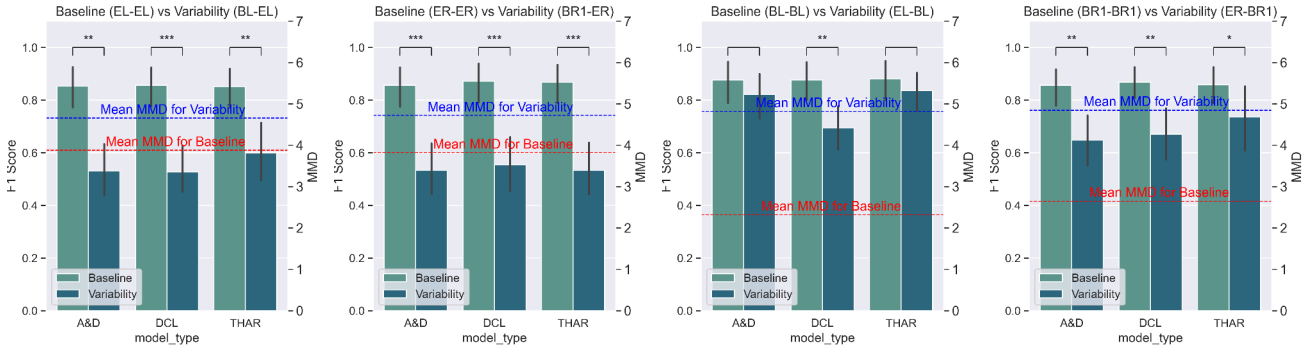
(a) EL - ER **(b)** ER - EL **(c)** BL - BR1 **(d)** BR1 - BL

Figure 5.2: Performance changes due to Positional Variability. Bars represent the average F1 score for each DL HAR model, and the lines represent the average MMD values of the settings. Light green bars represent the no variability setting; dark green bars represent the setting with variability. Asterisks represent the p-value of a paired t-test (*: $p - value < 0.05$, **: $p - value < 0.01$, ***: $p - value < 0.001$). Significant performance changes were found for all models when BlueSense sensors were used but not for Empatica sensors.

differences between Bluesense and Empatica sensors and how positional variability causes a shift in their data distribution (as evidenced by MMD values).

We note that the baseline performance of the DL models (indicated in light green) is consistent and independent of the sensor used, as shown throughout the experiments in Figure 5.2.

The drop in performance due to position variability is inconsistent across models when empatica sensors are used. In Figure 5.2a, we only see DeepConVLSTM show a significant drop in performance ($p < 0.001$), whereas in Figure 5.2b, Attend&Discriminate is the only model with a significant drop in performance ($p < 0.05$). From the experiments done using the empatica sensors, we see that DL models, in general, can be robust against positional variability for simple activities such as walking. When the device used is Bluesense (Figures 5.2c and 5.2d), all DL models experience a significant drop in performance, resulting in a mean F1 score between 0.4 and 0.45—essentially equivalent to random guessing.



(a) BL - EL **(b)** BR1 - ER **(c)** EL- BL **(d)** ER - BR1

Figure 5.3: Performance changes due to Device Variability. Bars represent the average F1 score for each DL HAR model, and the lines represent the setting's average MMD values. Light green bars represent the no variability setting; dark green bars represent the setting with variability. Asterisks represent the p-value of a paired t-test (*: $p - value < 0.05$, **: $p - value < 0.01$, ***: $p - value < 0.001$). Significant differences in the performance were found in all but two cases.

5.1.3 Device Variability

Device variability, shown in Figure 5.3, caused the most significant performance drop (p-value ≤ 0.001 for most cases) in the three DL HAR models compared to Position and Orientation Variability. The Device Variability experiments can be subdivided into two categories:

1. Train bluesense and Test Empatica. Figures 5.3a and 5.3b

2. Train Empatica and Test Bluesense. Figures 5.3c and 5.3d.

We see that the performance drop due to Device Variability is larger for 'Train bluesense and Test Empatica' scenarios (mean F1-Score drop of 0.35) vs 'Train Empatica and Test Bluesense' scenarios (mean F1-Score drop of 0.17). Significant changes in DL model performance are observed in all three models in category 1.

5.1.4 Subject Variability

Subject variability becomes evident when we examine the results at a granular level. Instead of just focusing on the mean performance, looking at each leave-one-subject-out cross-validation (CV)

result reveals that the F1 scores for individual participants vary significantly for both baseline and variability scenarios. Figure 5.4 illustrates an example of position variability by comparing sensors worn on the left and right wrists (Empatica-Left and Empatica-Right), detailing the F1 score per participant in ascending order of baseline F1 scores. Out of 16 participants, the first six deviate from the average trend. Participants 9, 4, and 2 exhibit very poor F1 scores of 0.4, indicating that the model's performance was comparable to making random guesses. Participants 3, 5, and 1 performed better in the variability scenario than in the baseline scenario. All other participants followed the mean trend, where baseline performance was higher than performance in the variability setting.

With these results, we can observe how **variability reveals model performance nuances**. Throughout all the experiments conducted, we observed consistent mean performance across all models in the baseline experiments. These baseline experiments mimic the typical testing conditions for DL HAR models. Without variability, all DL-HAR models exhibited similar high performance. However, when we isolated a specific type of variability in our experiments, we observed varying performance among the different DL models, with some models exhibiting larger performance drops than others. These differences were inconsistent across experiments, with some models having larger differences in one experiment and smaller in another. Evaluating models with variability reveals their nuances and behavior under real-life conditions, demonstrating how they adapt to such changes. These results highlight the importance of testing DL HAR models under realistic conditions to better understand their robustness and adaptability.

5.2 Understanding Model Performance with MMD Metric

As observed in the previous section, the effect of variability in model performance is unequal across types of variability, models, subjects and the selected test sensor. We hypothesized that the performance drop is related to the ”amount of shift” in data distribution induced by the variability. To validate this, we use the Maximum Mean Discrepancy (MMD) metric, measuring the distance between two distributions.

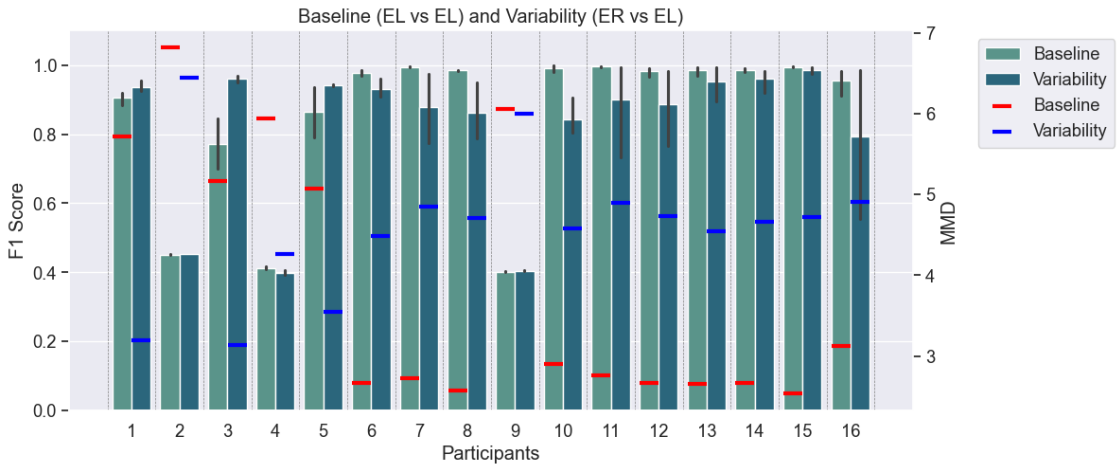


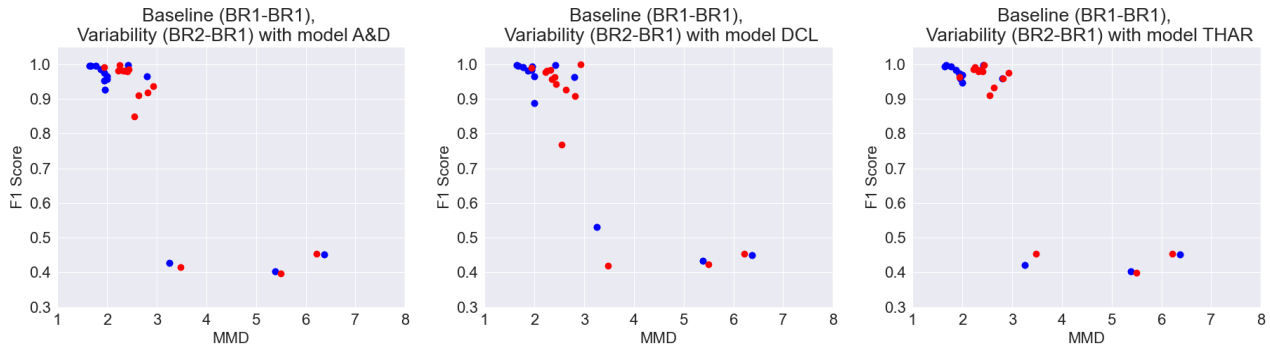
Figure 5.4: MMD of train vs test data and its relationship to the average F1-Score of the three evaluated models in the ER-EL experiment. This example depicts position variability where EL sensor data is used for testing. Dark green bars represent the F1-Score under variability, light green represents the baseline F1-Scores, and the blue and red points are their respective MMD values. The CV are arranged in ascending order of baseline F1 score.

Figures 5.1, 5.2, and 5.3 depict the average MMD for the baseline setting with a red line, and the average MMD for the variability setting with a blue line. MMD is the same for all DL models in a given setting, as the train and test set are the same in each experiment. Since a higher MMD value indicates a greater shift in data distribution, higher MMD values represent more dissimilar distributions between the test and train data.

5.2.1 MMD to Explain Orientation, Position and Device Variability

We first study **differences in performance due to each type of variability**. Observing the average MMD values for all the experiments, it is apparent that the MMD is lower for the baseline than for the variability setting. This supports the hypothesis that variability causes a shift in data distribution. Moreover, the difference in MMD between the two settings is related to the difference in F1-Score, supporting the hypothesis that MMD is correlated with performance.

In Figure 5.1, the small difference between baseline and variability MMD values aligns with the minimal performance drop observed due to orientation variability. Examining each cross-validation in Figure 5.5 reveals that both variability scenarios (in red) and baseline scenarios (in



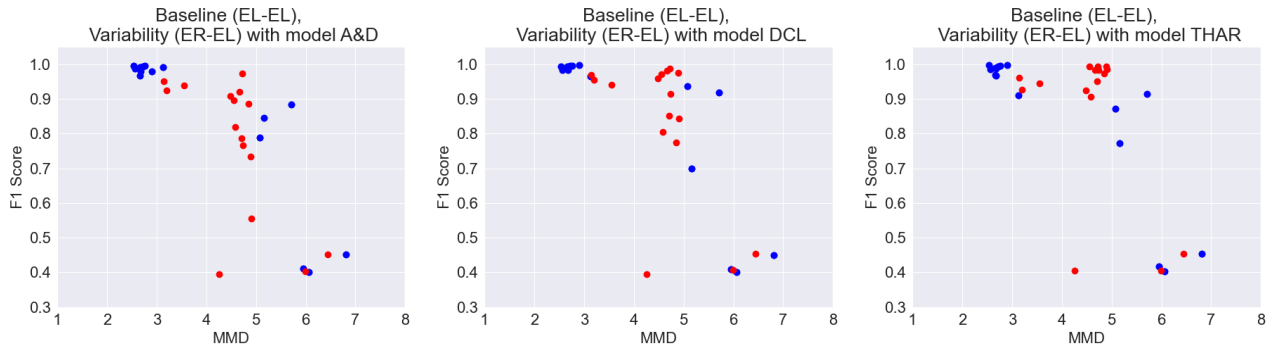
(a) A&D **(b)** DCL **(c)** THAR

Figure 5.5: Variability has little effect on the value of MMD or F1 score when observing effects of Orientation Variability. Here, when depicting a variability Scenario as BR2-BR1, means that the model is trained with sensor BR2 and tested with sensor BR1

blue) exhibit comparable MMD and F1 scores, with red points showing only slightly higher MMD values. This trend, consistent across all three models studied, suggests that DL models can demonstrate robustness to variability up to a certain threshold of data distribution shift.

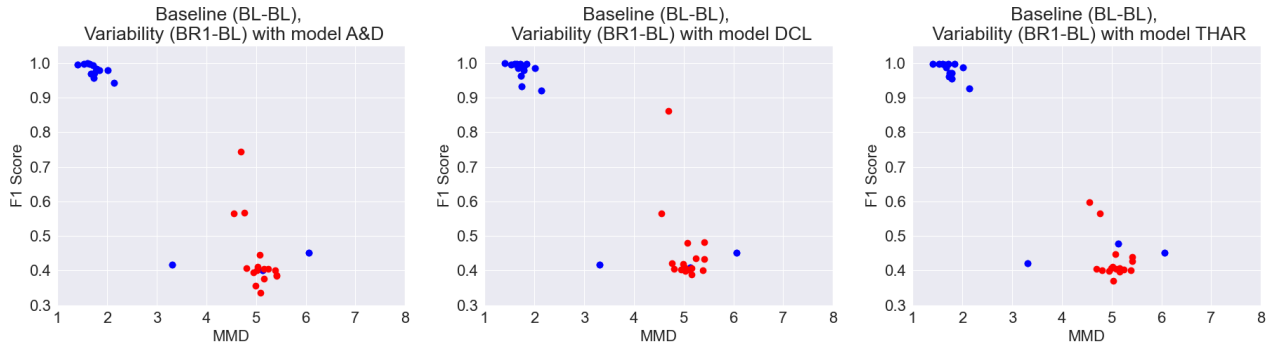
Observing positional variability, in Figures 5.2a and 5.2b, a small difference in mean MMD values aligns with a small drop in F1-Score. In contrast, in Figures 5.2c and 5.2d, a greater difference in MMD values corresponds to a significant drop in DL model performance. These observations indicate a relationship between MMD values and the performance change in DL HAR models due to variability. The MMD difference can explain the greater performance drop when Bluesense sensors are used as a test sensor compared to Empatica sensors in position variability scenarios.

In Figures 5.6 and 5.7, we observe that while the MMD values due to positional variability (red points) are similar for both Bluesense and Empatica sensors (approximately 5), the Empatica sensors generally achieve higher F1 scores on average. Additionally, the negative relationship between MMD and F1 score does not follow a simple linear trend with a negative slope as suggested by [37]; instead, there appears to be an MMD threshold beyond which model performance begins to decline. This suggests that MMD alone cannot fully account for performance drops due to variability but rather serves as an indicator of distribution shift—one of multiple factors possibly influenced by variability.



(a) A&D **(b)** DCL **(c)** THAR

Figure 5.6: Relationship between the F1 score and the MMD values under Position Variability when Empatica sensors (ER and EL) are used. The drop in performance starts to happen when MMD values are between 5 and 6.

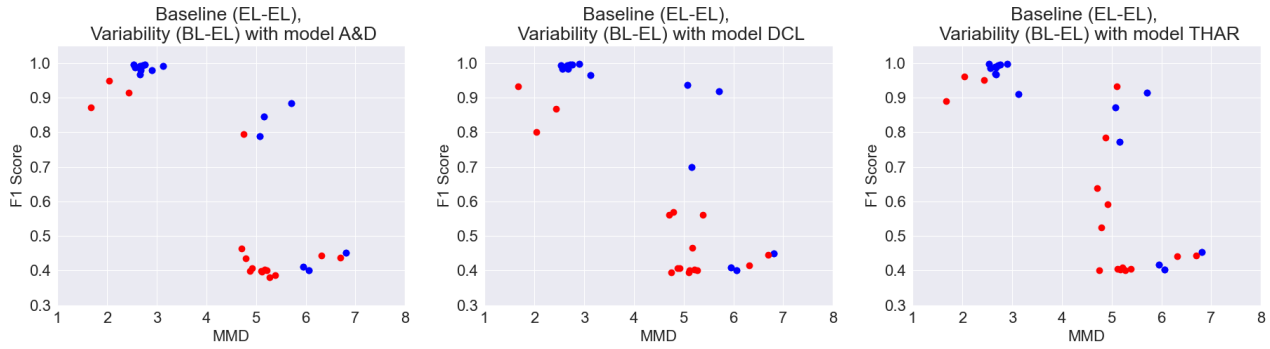


(a) A&D **(b)** DCL **(c)** THAR

Figure 5.7: A significant drop in performance as MMD value approaches 5 when observing Position Variability using Bluesense sensors.

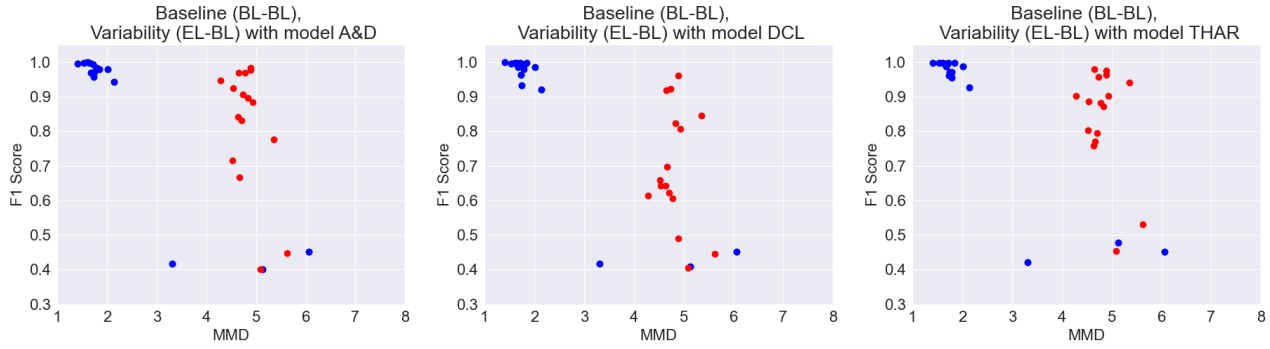
In Figure 5.7, the baseline MMD value (blue dots) for Bluesense sensors is around 1.5, which is significantly lower than that of the baseline MMD for Empatica sensors in Figure 5.6. Conversely, MMD values under variability conditions for Empatica and Bluesense sensors hover around 5. This suggests that Bluesense sensors capture data at a higher sampling rate and have a lower distribution shift between participants. In contrast, the higher baseline MMD for Empatica implies a more inherent distribution difference between participants. This greater inter-participant variability captured by Empatica sensors seems to improve model robustness to positional variability.

Figure 5.6 representing Empatica sensors shows that for similar MMD values, models yield higher F1 scores than with Bluesense sensors as shown in Figure 5.7.



(a) A&D **(b)** DCL **(c)** THAR

Figure 5.8: When a model is tested using Empatica sensors and trained using Bluesense to highlight Device Variability, the model performs poorly.



(a) A&D **(b)** DCL **(c)** THAR

Figure 5.9: When a model is tested using Bluesense sensors and trained with Empatica sensors in the Device Variability scenario (Red points), even after a higher MMD value between 4 and 5, we see only a significant drop in performance when using the DeepConvLSTM model.

Figure 5.3 presents a contrasting outcome to the positional variability observations in Figure 5.2. Here, instead of a proportional drop in performance relative to the difference in mean MMD values, we observe that a smaller MMD difference is related to a larger performance drop in Figures 5.3a and 5.3b. Conversely, in Figures 5.3c and 5.3d, the MMD difference is larger, but the performance drop is smaller and less significant.

In Figure 5.8, when models trained on Bluesense sensors are tested on Empatica sensors (red points), the F1 score clusters around 0.4 when MMD is approximately 5, indicating poor classification performance. By contrast, Figure 5.9 shows that when trained on Empatica sensors and tested on Bluesense sensors, F1 scores vary more broadly between 0.6 and 1 at a similar MMD value of 5, with TinyHAR achieving scores between 0.7 and 1. This suggests that models generalize better when trained on Empatica sensors and tested on Bluesense rather than vice versa. As observed in positional variability analysis, the MMD-F1 score correlation here is not a simple linear decline. Instead, MMD values around 5 act as a threshold, with F1 scores close to this threshold showing variable performance. Scores for MMD greater than 5 remain around 0.4, while scores below 5 generally range from 0.8 to 0.9.

This discrepancy in performance for the same MMD can be attributed to the differences in sampling rates between the sensors used in the experiments. Bluesense sensors sample at 100Hz, while Empatica sensors sample at 64Hz. This means that for the same 2-second time window, Bluesense sensors provide 200 samples, whereas Empatica sensors provide 128 samples. When a model is trained with Bluesense data (higher sampling rate) and tested with Empatica data (lower sampling rate), we must upsample the Empatica data. Upsampling does not introduce higher frequency features into the data, which might be essential for the model's accurate classification if trained with higher frequency information. On the other hand, if a model is trained with Empatica data and tested with Bluesense data, we downsample the Bluesense data. Downsampling removes high-frequency features from the test data, which the model, trained on lower-frequency data, does not rely on. Therefore, the performance drop is not as significant.

Another factor to consider is model complexity. Models trained with Bluesense sensors take longer inputs for the same time window than those trained with Empatica sensors, which means that the number of inputs in each layer is larger (Table 4.2). More Complex models may become highly specialized to the training data, which can increase their susceptibility to variability. This is because their complexity allows them to capture subtle details in the training data, which may not generalize well to data with different characteristics, leading to decreased performance when faced with variability. In contrast, less complex models might generalize better and thus perform more

consistently under variability conditions. Further tests are required to confirm this and to explore whether more data can make the models more robust to variability. Nonetheless, it is important to remember that the amount of labeled sensor data available for HAR is usually small.

5.2.2 MMD to Explain Subject Variability

We investigated the **differences in performance across participants** to highlight subject variability. We observed a high standard deviation in the F1-Score for each model, implying that each participant's performance depends on the participant's activity characteristics. To evaluate this, we measured the MMD for each cross-validation fold. For example, in experiment 2, position variability, as shown in Figure 5.4.

For participants 9, 4, and 2, who achieved F1 scores around 0.4 (indicating the model struggled to distinguish between walking and not walking), their MMD values were notably higher. This aligns with the fact that these participants held onto support bars during the treadmill experiment, as shown in Table 3.2, highlighting how slight variations in activity execution can heavily impact model performance.

Moreover, participants 3, 5, and 1 present an exception: their baseline MMD is higher than in the variability scenario. These participants performed better in the variability scenario but worse in the baseline scenario, which suggests that their test data in the variability setting was more similar to the training data compared to the baseline.

A consistent pattern emerges for participants 16, 12, 6, 13, 14, 10, 8, 15, 7, and 11: low MMD in the baseline setting and high MMD in the variability setting. This explains their higher performance in the baseline scenario and the drop in performance when variability was introduced.

5.2.3 MMD Correlation to F1 Score

Upon calculating the correlation between the F1 score and the MMD between the train and test sets, we observed a negative correlation, as illustrated in Figure 5.10. This supports our hypothesis that a relationship exists between the shift in data distribution and model performance. Almost

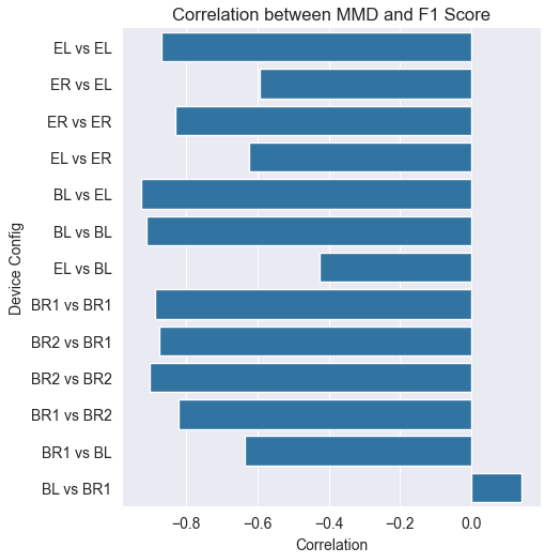


Figure 5.10: The correlation of the MMD values between train and test to the F1 score is mostly negative. Showing there is a negative correlation between the MMD and the Performance of a DL model.

all experiments demonstrated this negative correlation between the F1 score and MMD, further validating our hypothesis.

However, an exception was found in the Bluesense-Left (BL) vs. Bluesense-Right (BR1) sensor experiment, where the correlation was closer to 0. This outlier can be attributed to the consistently poor performance of the models across all participants in the cross-validation, regardless of the MMD value. In scenarios where the model performs poorly overall, the impact of changes in MMD appears minimal.

We observed how introducing variability, whether from orientation, position, device, or subject, results in higher MMD in most cases. The MMD has shown how, for some participants, introducing variability helps the data become more similar to the train distribution, explaining why, in some cases, the performance increases when variability is introduced. This relationship underscores the impact of data distribution shifts on the performance of DL HAR models and highlights the importance of considering individual participant variability in model evaluation.

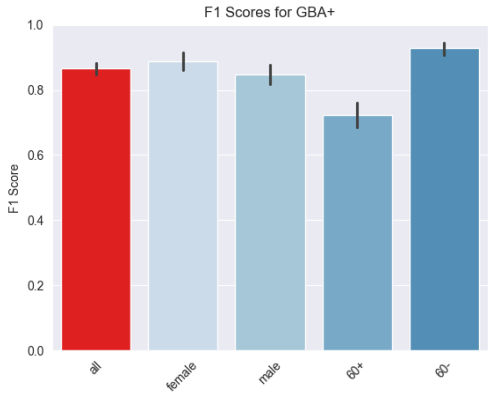


Figure 5.11: Average Baseline F1 score for all experiments for various demographics.

5.3 Gender-Based Analysis using HARVAR Results

This GBA+ (Gender-Based Analysis Plus) examines potential differences in model performance across demographic groups under 'variability' conditions. Participants were categorized by self-reported sex (female and male) and age (under 60 and over 60 years), and model performance for each group was compared to the overall baseline (marked as 'all'), represented by the mean F1 score from previous sections. Given the limited sample size of 16 participants, these findings offer preliminary insights and should be interpreted cautiously.

5.3.1 GBA+ of Orientation Variability on DL HAR Models

Figure 5.11 illustrates the baseline performance discrepancies of DL HAR models across demographics. Results indicate a 0.03 higher performance when participants are female than male. The model performance differences between male and female participants are not attributable to data imbalance, given the relatively even split (7 female, 9 male). Models demonstrate a slight performance advantage when evaluated on data from female participants, even when the training data has a majority of males.

The lowest model performance is observed in the group aged 60 and over, with a notable drop of 0.15 in F1 score compared to the "all" category (representing the mean F1 score). This drop suggests inherent differences in data characteristics between participants over 60 and those under

60. Additionally, the dataset's composition—5 participants aged 60+ versus 11 participants under 60—may contribute to this disparity, as the models likely favor the majority (younger) training data.

Figure 5.12 demonstrates that the effect of orientation variability is largely consistent across all demographics, with a notable exception in the 60+ age group, where the impact is minimal (approximately 0.02). When comparing mean F1 score drops across gender groups, the difference is modest: female participants show a 0.04 reduction, whereas male participants exhibit a more pronounced drop of 0.1. These findings suggest a slightly greater sensitivity to orientation variability in male participants, while participants over 60 appear comparatively unaffected by orientation shifts.

In analyzing the effects of positional variability across demographics as illustrated in Figure 5.13, significant differences emerge between sensor types. With Bluesense sensors (Figure 5.13a), the effect of positional variability is detrimental, causing F1 scores to drop to a range between 0.4 and 0.5. Such low performance is comparable to random guessing, indicating a significant degradation in model reliability across all demographics.

Conversely, with Empatica sensors (Figure 5.13b), the impact of positional variability is minimal across all groups. Notably, there is no performance drop in the 60+ demographic; rather, there

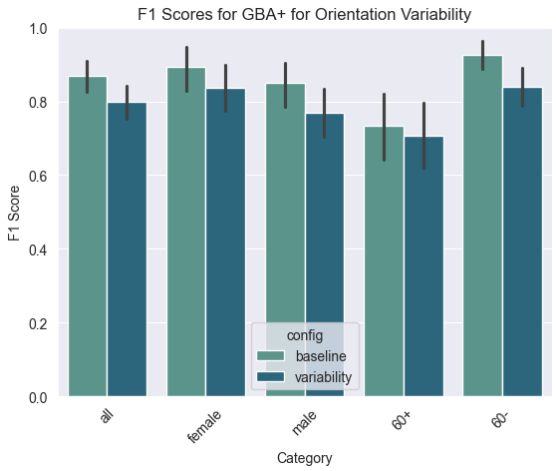
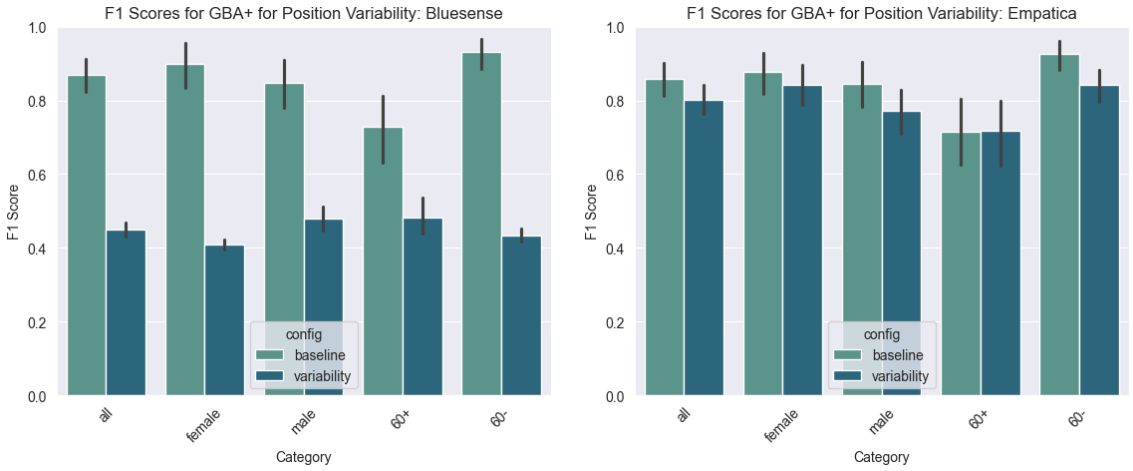


Figure 5.12: Effect of Orientation Variability on the Performance of DL HAR models when tested with participants of various demographics in GBA+.



(a) Using Bluesense Sensors **(b)** Using Empatica Sensors

Figure 5.13: GBA+ analysis conducted to assess the impact of positional variability on DL HAR model performance. Positional variability scenarios were divided into two categories based on the sensors used: Bluesense and Empatica.

is a slight performance increase. The F1 score reductions across other demographics—0.03 for females, 0.05 for males, and 0.08 for participants under 60—are modest and comparable to the variability impacts observed in orientation variability tests.

(a) Using Bluesense Sensors for testing **(b)** Using Empatica Sensors for testing

Figure 5.14: GBA+ analysis conducted to assess the impact of device variability on DL HAR model performance. Device variability scenarios were divided into two categories based on the sensors used for testing: Bluesense and Empatica.

The device variability GBA+ analysis categorizes results based on the sensors used for testing. In Figure 5.14a, where Bluesense sensors are used for testing, baseline models were trained and tested on Bluesense, while models under variability were trained on Empatica and tested on Bluesense. Meanwhile, in Figure 5.14b, Empatica sensors are used for testing, with baseline models trained and tested on Empatica and variability models trained on Bluesense and tested on Empatica.

When Bluesense sensors are the test devices (Figure 5.14a), female participants' test data shows a higher drop from device variability, with an F1 score drop nearing 0.2, while male participants experience a smaller drop of about 0.1. Additionally, participants under 60 experience a larger decrease of 0.2, whereas the 60+ group's performance declines by only 0.1. Relative to the average performance drop of 0.15 across the entire dataset (represented by the "all" category), the male and 60+ groups appear less affected by device variability.

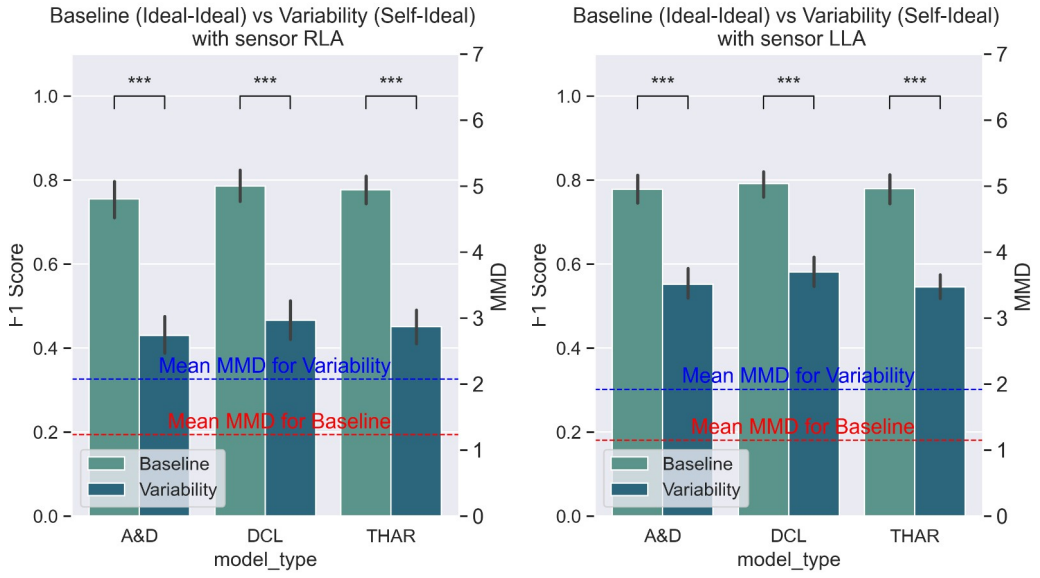
On the other hand, Figure 5.14b shows the effects of device variability when Empatica sensors are used for testing. Here, there is a substantial drop in performance across all demographic groups, except for the 60+ group, which experiences only a minor decline of 0.02 in F1 score. In contrast, all other groups see a performance drop greater than 0.3, highlighting the substantial impact of device variability on model accuracy when switching from Bluesense to Empatica.

5.4 Compounding Variability Effects in Real-Life Scenarios (REALDISP Case Study)

The results from the REALDISP dataset revealed a significant drop in performance for both RLA and LLA sensors due to the compounding effects of variability (p-value < 0.001), as shown in Figure 5.15. Figure 5.15a illustrates the performance of DL models trained on data collected from the RLA sensor. Consistent with the findings from the HARVAR dataset, a higher MMD value corresponds to scenarios with poorer performance, while a lower MMD value corresponds to scenarios with better performance. Specifically, the MMD between the Ideal train and test data is much lower than the MMD between the Self-train and Ideal test data.

When analyzing the performance using the LLA sensor in figure 5.15b, we observe that in the Ideal vs. Ideal scenario, DL models perform similarly regardless of whether the RLA (F1 score 0.76) or LLA (F1 score 0.78) sensor data is used. However, in the Self vs. Ideal scenario, the LLA sensor outperforms the RLA sensor. The mean F1 score for the Self vs. Ideal scenario is 0.55 when using the LLA sensor, compared to 0.44 with the RLA sensor. This difference in performance is reflected in the MMD values: the MMD for LLA-Self vs. LLA-Ideal is 1.9, while RLA-Self vs. RLA-Ideal has an MMD of 2.05. These results further confirm that a lower MMD value corresponds to better model performance, while a higher MMD value indicates worse performance.

Figure 5.16 shows the mean F1 score and MMD values for each scenario outlined in Table 4.3 for both RLA and LLA sensors. The best performance is observed in the scenario where both the training and testing data are collected under ideal conditions, which is expected since there is no variability to degrade the performance of the DL model. The poorest performance occurs when the



(a) RLA Ideal - RLA Self **(b)** LLA Ideal - LLA Self

Figure 5.15: The mean F1 score and MMD values for the experiments conducted using the REALDISP dataset. The mean values are calculated over all the cross-validation folds. Asterisks represent the p-value of a paired t-test (*: $p - value < 0.05$, **: $p - value < 0.01$, ***: $p - value < 0.001$). The effect of variability on the model performance was greater when the right wrist sensor (RLA) was used.

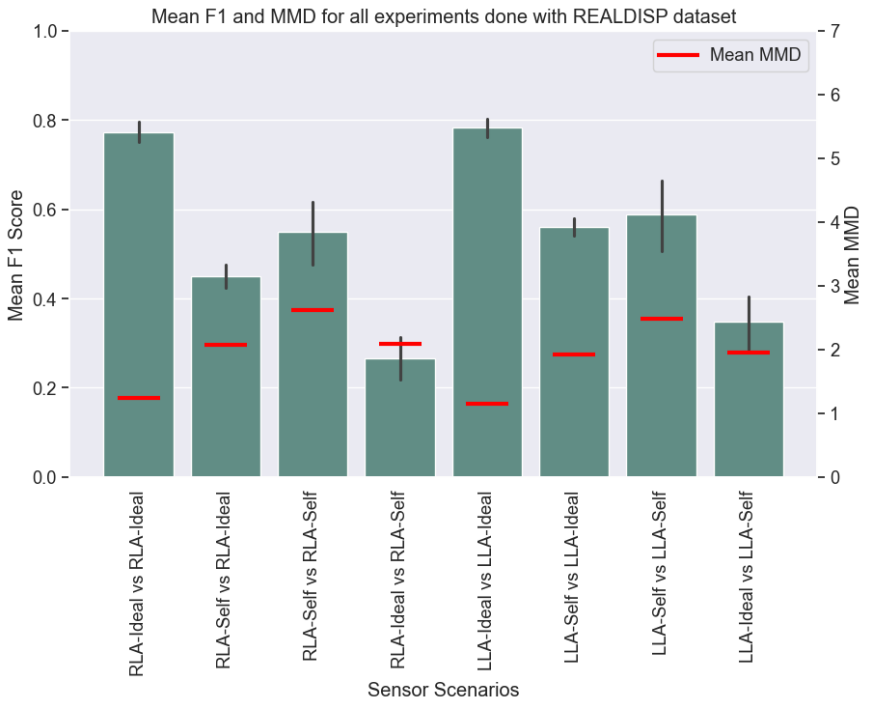


Figure 5.16: The mean F1 score and MMD values for all scenarios tested using the REALDISP dataset.

model is trained on ideal data but tested on self-collected data. This indicates that any DL HAR model trained using lab-collected data would likely perform poorly when applied in real-world settings with position and orientation variabilities.

The MMD values reveal that the highest MMD occurs when training and testing data are of type self. While we have observed that higher MMD values generally correspond to lower performance, this trend does not hold in this case. The elevated MMD in the self vs. self scenario can be explained by the significant variability within the self data, as participants wear sensors in varied ways, often with sensors flipped across axes. This variability leads to a wider distribution shift, resulting in a higher MMD value. However, this diversity in the training data makes the model more robust to variability, leading to better generalization and performance in the self vs. self scenario.

In contrast, the ideal vs. self scenario suffers because the models trained on ideal data lack exposure to variability during training, making them vulnerable when tested under non-ideal conditions. Notably, despite having similar MMD values to the ideal vs. self scenario, the self vs.

ideal scenario performs better. This can be attributed to the fact that when a DL HAR model is trained on diverse and variable data, it becomes more robust, resulting in improved performance even when tested on ideal data.

Chapter 6

Conclusion

This chapter discusses the implications of the findings of this study in Section 6.1. Next, we address the limitations of this study and the HARVAR dataset in Section 6.2. Finally, we conclude with Summary and Future Works in Section 6.3

6.1 Study Implications

This section discusses the key implications of our study's findings, specifically how different types of variability impact DL HAR model performance and what these effects mean for real-world applications. We discuss distribution shifts caused by variability and give insights into how these shifts influence model performance.

6.1.1 Position and Orientation Variability Implications on Real-World Scenarios

Across the three types of variability studied in this paper, orientation variability caused the lowest performance drops across all models. This result suggests that models trained on IMU data from devices worn in fixed positions, such as smart glasses and earbuds, have more chances to generalize to multiple participants and environments, as the orientation variations that may occur will not significantly impact performance.

Smartwatches are particularly vulnerable to the compounded effects of orientation and position variabilities. Experiments with the REALDISP dataset highlighted the significant impact of these variabilities on wrist-worn sensors, where orientation changes can be as extreme as a 180-degree flip across an axis. This drastic orientation shift further amplifies the effect on model performance when combined with position variability. Our findings also indicate that training models with a diverse dataset that includes a range of variabilities results in more robust performance, making them better suited for real-world applications.

6.1.2 Device Variability has a High Impact on DL Model Performance.

Device variability drastically impacted the performance of DL models because it not only causes a shift in data distribution but also introduces differences in sampling frequency. These changes can affect the model size and necessitate resampling when using a device different from the one used in training. In addition, for this type of variability, MMD is insufficient to understand the variability.

Given the evolving wearable device industry, device variability is one of the main challenges to build truly generalizable HAR models. Currently, different models for each device are required, which means updating models every time, which can be prohibitive if no data for the device has been collected. Researchers have utilized fine-tuning and domain adaptation methods in [28] to overcome the effect of device variability in cross-dataset scenarios, where one dataset is used to train a model and another is used to test it.

Enhancing model robustness is crucial to address device variability, but it also requires careful data preprocessing and determining the optimal sampling rate for training the model. This would ensure that the model can generalize better across different devices.

6.1.3 Subject Variability and the Need for Diverse Training Data

The HARVAR results showed that variability in how individuals perform activities significantly impacts the performance of DL models. Human activity is inherently variable; these differences can change with age, demographics, and personal preferences. In our study, participants were

asked to perform a simple treadmill walking task without specific instructions, leading some to hold the side bars while others did not. The reduced movement caused by holding the sidebars made it difficult for the models to classify walking accurately for those participants.

However, it is important to note that the training data for the DL models was not completely isolated from sidebar holding, as two participants in the training set also held onto the bars. Despite this, for CVs 9, 4, and 2, the sidebar holding data in train sets was outnumbered by non-side bar holding data in a ratio of 2:13. This highlights the models' bias towards the majority of the training data. To improve performance and generalizability for larger populations, datasets must either ensure better balance across activity variations or apply preprocessing techniques that give more weight to underrepresented data in the training set.

6.1.4 Larger MMD Correlate with Smaller F1-Score, with Limitations

MMD serves as a useful metric to calculate the shift in data distribution. We observed a strong correlation between MMD and F1-Score, such as when MMD is large, F1-Score is low, and vice versa. Still, it sometimes fails to fully capture the impact of variability, as observed in the case of device variability. When changing devices alters the input shape to a model, MMD may not adequately explain the variability.

Additionally, MMD is a better metric when the F1 score is high, i.e., when the model's performance is good. However, beyond a certain threshold, when MMD is too high, changes in MMD stop reflecting in the changes to performance. As seen in Figure 5.4, spikes in MMD values for participants 2, 4, and 9 vary, but these three participants show an average F1 score of 0.41. On the other hand, when the performance is high, the difference in MMD shows a clear inverse relationship.

6.1.5 No Significant Differences in Performance Change Across the Three Models

Statistical tests revealed no significant difference in performance between the three evaluated models. However, models with larger MACs tend to have bigger performance drops. High model complexity results from a larger input size due to a higher sampling rate or a larger network with more layers. We assume that increased complexity allows models to learn finer features, making them more prone to overfitting and less adaptable to changes. This aligns with previous results, such as the shallowLSTM [11] network, which showed that using one less layer in the DCL model results in higher performance. In the face of non-significant performance changes, we recommend using lighter models, such as the TinyHAR model, which achieves similar performance and robustness with fewer parameters.

6.2 Study Limitations

This study isolated the effects of each type of variability in a binary classification task, while the DL models are capable of multiclass classification, as shown using the REALDISP dataset. Further studies are needed using multiclass classification with diverse activities in terms of motion and duration to better understand the robustness of the models.

We evaluated the effects of variability in two datasets, each with 16 participants. While this number is small but similar to other public HAR datasets. However, the small size might not be enough to reveal significant differences across models and for some experiments. Increasing the number of participants can help reveal differences across the models, but larger datasets do not showcase the same type of variabilities observed in these two datasets.

We studied wrist-worn sensor variabilities (orientation, position, and device). Future research should consider variability in other sensor placements, such as earbuds, chest-mounted sensors, and smart glasses. Device variability was only tested between two devices with 64Hz and 100Hz sampling frequencies, while many other devices with different noise levels and sensitivity ranges exist. As this type of variability showed the highest drops in performance, a deeper study on its

effects and how to overcome it might be required. Other research [26] have also found that the sampling rate of the train and test data should match for optimal performance.

Finally, we focused solely on accelerometer data, whereas many DL models are designed to combine multiple modalities, including gyroscope and magnetometer data, for HAR. We used only the accelerometer as it is the most common modality in many devices and has the lowest power consumption, making it preferable when possible.

6.3 Summary and Future Work

In this work, we have studied three types of variability in three different DL-HAR models. We isolated each type of variability in our experiments, done with the HARVAR dataset specifically collected for this study. We evaluated the distribution and performance changes caused by position, orientation, and device changes.

Our findings reveal that different types of variability—orientation, position, and device—affect DL HAR models in distinct ways, with device and position variability causing the largest performance drops, particularly when using certain sensors like Bluesense. We observed a strong correlation between distribution shifts (measured by MMD) and model performance declines, suggesting that MMD can serve as a useful predictor for performance drops when models are deployed in varied settings. Additionally, high-complexity models and those trained with high sampling rates performed worse under variability, likely due to reliance on fine-grained features that don't generalize well across contexts. Although our study was limited to a single activity for isolated variability analysis and a small participant pool, the combined results using the HARVAR and REALDISP datasets suggest that integrating diverse variability into training data can improve model robustness. The MMD metric and train-test pipeline offer tools for assessing model robustness to variability in future DL HAR research.

In future works, we should evaluate DL HAR models by focusing on their performance in detecting beyond simple and periodic activities like walking. The activities from the 'Salad Preparation' segment of the HARVAR dataset that were used as 'not walking' in our study can be labeled

and divided broadly into two categories of sub-activities: gestures (like biting or sipping) and complex activities (like washing dishes or seasoning the salad). Conducting a similar study with a wide spectrum of activity types would allow a complete understanding of the change in model performance due to distribution shifts.

In addition, the methodology applied in this study—exploring the impact of subject, device, position, and orientation variability on model performance—can be extended to other datasets beyond HARVAR and REALDISP. Although the existing datasets might not be able to perform an isolated study on all types of variabilities under a baseline and variability setting, they can follow an experiment process similar to the one conducted using the REALDISP dataset. Either mixed effects of variabilities can be studied, or only one variability can be studied. For example, the OPPORTUNITY dataset can be used to study the effect of subject and position variability. It uses multiple sensors worn at different body positions on several different participants. Validating results from other datasets would help strengthen the implications of this study, and contradicting results will allow more advanced insight.

Future studies can use MMD as a metric to quantify the diversity of datasets by comparing the distribution of various participants. This approach can help determine whether using diverse datasets to build a generalized model is more effective or if customized models tailored to specific demographics or personalized models provide better performance.

In this study, we found that device variability significantly impacts the performance of DL models, with nuanced effects and highly dependent on the specific devices involved. Future work could focus on developing preprocessing techniques to improve model transferability across different sensors. By identifying the most effective ways to adapt a model trained on one device for use with another, these studies could enhance model robustness and broaden the practical applicability of DL HAR models across diverse sensor types.

MMD's capability to quantify distribution differences can be studied in unsupervised HAR applications, where the availability of labeled data is often limited. In these contexts, MMD could be applied to cluster unlabelled data, grouping similar activities based on their distributions to enable unsupervised classification.

In conclusion, this study sets the stage for a broader exploration of the effects of real-world variabilities on DL HAR models. By understanding, quantifying, and addressing these effects, future work can contribute to developing HAR models that are both robust and adaptable, supporting reliable performance across a wide range of applications and demographics.

List of Figures

3.1 Placement of sensors in HARVAR data collection. The Empatica Embrace Plus and Bluesense sensors are placed in the same coordinate system, and their axis is marked. BR2, marked as red, is tilted across the Z-axis at 45-degree of rotation. In this diagram, the person is facing towards the reader 18

3.2 Screenshots of the data annotation application used to roughly annotate the activities. The application was used by researchers who observed the activities done by the participants. 19

3.3 The figure on the left depicts the Bluesense-LWR and the Empatica-Left sensor magnitude before their timestamps were synchronized. On the right is the depiction of signals after synchronization has been performed. This data is of Participant 7 and shows the clap that was performed before the walking activity. 21

3.4 The file structure in which the HARVAR Data is stored. In this figure, pxxx can be any participant number, such as p001 to p016. 22

4.1 The experiment setting using the HARVAR dataset to evaluate the effect of device, position, and orientation variability. Where Sensor 1 and Sensor 2 are used in combination as a train-test pair to highlight variability. In these diagrams, the person is facing towards the reader. 26

4.2 The process of evaluating the effect of variability using the HARVAR dataset. . . . 28

4.3 The process of calculating the MMD between the train and test data of the HARVAR dataset. 31

5.1 Performance changes due to Orientation Variability. We show the average F1 score and average MMD values for each DL HAR model in the two experiments. Light green bars represent the no variability setting of each experiment, and dark green bars represent the variability setting. Asterisks represent the p-value of a paired t-test (*: $p-value < 0.05$, **: $p-value < 0.01$, ***: $p-value < 0.001$). Only two models in one experiment showed significant performance changes, but the F1-Score remains above 0.7. 35

5.2 Performance changes due to Positional Variability. Bars represent the average F1 score for each DL HAR model, and the lines represent the average MMD values of the settings. Light green bars represent the no variability setting; dark green bars represent the setting with variability. Asterisks represent the p-value of a paired t-test (*: $p-value < 0.05$, **: $p-value < 0.01$, ***: $p-value < 0.001$). Significant performance changes were found for all models when BlueSense sensors were used but not for Empatica sensors. 37

5.3 Performance changes due to Device Variability. Bars represent the average F1 score for each DL HAR model, and the lines represent the setting's average MMD values. Light green bars represent the no variability setting; dark green bars represent the setting with variability. Asterisks represent the p-value of a paired t-test (*: $p-value < 0.05$, **: $p-value < 0.01$, ***: $p-value < 0.001$). Significant differences in the performance were found in all but two cases. 38

5.4 MMD of train vs test data and its relationship to the average F1-Score of the three evaluated models in the ER-EL experiment. This example depicts position variability where EL sensor data is used for testing. Dark green bars represent the F1-Score under variability, light green represents the baseline F1-Scores, and the blue and red points are their respective MMD values. The CV are arranged in ascending order of baseline F1 score. 40

5.5 Variability has little effect on the value of MMD or F1 score when observing effects of Orientation Variability. Here, when depicting a variability Scenario as BR2-BR1, means that the model is trained with sensor BR2 and tested with sensor BR1 . 41
5.6 Relationship between the F1 score and the MMD values under Position Variability when Empatica sensors (ER and EL) are used. The drop in performance starts to happen when MMD values are between 5 and 6. 42
5.7 A significant drop in performance as MMD value approaches 5 when observing Position Variability using Bluesense sensors. 42
5.8 When a model is tested using Empatica sensors and trained using Bluesense to highlight Device Variability, the model performs poorly. 43
5.9 When a model is tested using Bluesense sensors and trained with Empatica sensors in the Device Variability scenario (Red points), even after a higher MMD value between 4 and 5, we see only a significant drop in performance when using the DeepConvLSTM model. 43
5.10 The correlation of the MMD values between train and test to the F1 score is mostly negative. Showing there is a negative correlation between the MMD and the Performance of a DL model. 46
5.11 Average Baseline F1 score for all experiments for various demographics. 47
5.12 Effect of Orientation Variability on the Performance of DL HAR models when tested with participants of various demographics in GBA+. 48
5.13 GBA+ analysis conducted to assess the impact of positional variability on DL HAR model performance. Positional variability scenarios were divided into two categories based on the sensors used: Bluesense and Empatica. 49
5.14 GBA+ analysis conducted to assess the impact of device variability on DL HAR model performance. Device variability scenarios were divided into two categories based on the sensors used for testing: Bluesense and Empatica. 49

5.15 The mean F1 score and MMD values for the experiments conducted using the REALDISP dataset. The mean values are calculated over all the cross-validation folds. Asterisks represent the p-value of a paired t-test (*: $p - value < 0.05$, **: $p - value < 0.01$, ***: $p - value < 0.001$). The effect of variability on the model performance was greater when the right wrist sensor (RLA) was used. 51
5.16 The mean F1 score and MMD values for all scenarios tested using the REALDISP dataset. 52

List of Tables

2.1 The three SOTA DL HAR models used for this study and their key architectural differences. 8

3.1 The sensors used from the collection of the HARVAR dataset along with information on their sampling rate and placement. 18
3.2 Inforamtion about the 16 participants of HARVAR Dataset. 23
3.3 Duration of each labeled activity in HARVAR Dataset 23

4.1 List of Experiments conducted using the HARVAR Dataset. 27
4.2 The computational complexity (in MACs) variance between the different models. The computational complexity depends on the model architecture and the sensor being used to train due to the difference in sensor sampling rates. 29
4.3 Experiments to evaluate compound effects of variability using the REALDISP Dataset. 33

Bibliography

[1] Alireza Abedin, Mahsa Ehsanpour, Qinfeng Shi, Hamid Rezatofighi, and Damith C. Ranasinghe. Attend and discriminate: Beyond the state-of-the-art for human activity recognition using wearable sensors. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 5(1), mar 2021. https://doi.org/10.1145/3448083. URL https://doi.org/10.1145/3448083.

[2] Seong-Ho Ahn, Seeun Kim, and Dong-Hwa Jeong. Unsupervised domain adaptation for mitigating sensor variability and interspecies heterogeneity in animal activity recognition. *Animals*, 13(20), 2023. ISSN 2076-2615. https://doi.org/10.3390/ani13203276. URL https://www.mdpi.com/2076-2615/13/20/3276.

[3] Kolmogorov An. Sulla determinazione empirica di una legge didistribuzione. *Giorn Dell'inst Ital Degli Att*, 4:89–91, 1933.

[4] Ahmed Ayman, Omneya Attalah, and Heba Shaban. An efficient human activity recognition framework based on wearable imu wrist sensors. In *2019 IEEE International Conference on Imaging Systems and Techniques (IST)*, pages 1–5. IEEE, 2019.

[5] Mahsa Baktashmotlagh, Mehrtash Har, i, and Mathieu Salzmann. Distribution-matching embedding for visual domain adaptation. *Journal of Machine Learning Research*, 17(108): 1–30, 2016. URL http://jmlr.org/papers/v17/15-207.html.

[6] Oresti Baños, Miguel Damas, Héctor Pomares, Ignacio Rojas, Máté Attila Tóth, and Oliver Amft. A benchmark dataset to evaluate sensor displacement in activity recognition. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 1026–1035, 2012.

[7] Oresti Banos, Mate Attila Toth, Miguel Damas, Hector Pomares, and Ignacio Rojas.Dealing with the effects of sensor displacement in wearable activity recognition. *Sensors*, 14(6): 9995–10023, 2014.

[8] Billur Barshan and Murat Cihan Yüksek. Recognizing daily and sports activities in two open source machine learning environments using body-worn sensor units. *The Computer Journal*, 57(11):1649–1667, 2014.

[9] Henrik Blunck, Sourav Bhattacharya, Thor Prentow, Mikkel Kjrgaard, and Anind Dey. Heterogeneity Activity Recognition. UCI Machine Learning Repository, 2015. DOI: https://doi.org/10.24432/C5689X.

[10] Kjrgaard Mikkel Blunck Henrik, Bhattacharya Sourav Prentow Thor and Dey Anind. Heterogeneity Activity Recognition. UCI Machine Learning Repository, 2015. DOI: https://doi.org/10.24432/C5689X.

[11] Marius Bock, Alexander Hölzemann, Michael Moeller, and Kristof Van Laerhoven. Improving deep learning for har with shallow lstms. In *Proceedings of the 2021 ACM International Symposium on Wearable Computers*, ISWC '21, page 7–12, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384629. https://doi.org/10.1145/3460421.3480419. URL https://doi.org/10.1145/3460421.3480419.

[12] C Alan Boneau. The effects of violations of assumptions underlying the t test. *Psychological bulletin*, 57(1):49, 1960.

[13] Andreas Bulling, Ulf Blanke, and Bernt Schiele. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys (CSUR)*, 46(3):1–33, 2014.

[14] Ricardo Chavarriaga, Hesam Sagha, Alberto Calatroni, Sundara Tejaswi Digumarti, Gerhard Tröster, José del R Mill án, and Daniel Roggen. The opportunity challenge: A benchmark database for on-body sensor-based activity recognition. *Pattern Recognition Letters*, 34(15): 2033–2042, 2013.

[15] Sylvia Cho, Ipek Ensari, Chunhua Weng, Michael G Kahn, and Karthik Natarajan. Factors Affecting the Quality of Person-Generated Wearable Device Data and Associated Challenges: Rapid Systematic Review. *JMIR Mhealth Uhealth*, 9(3):e20738, Mar 2021. ISSN 2291-5222. https://doi.org/10.2196/20738.

[16] Rachel Danzig, Mengxi Wang, Amit Shah, and Lynn Marie Trotti.The wrist is not the brain: Estimation of sleep by clinical and consumer wearable actigraphy devices is impacted by multiple patient-and device-specific factors. *Journal of sleep research*, 29(1):e12926, 2020.

[17] Li Deng and John Platt. Ensemble deep learning for speech recognition. In *Proc. Interspeech*, September 2014. URL https://www.microsoft.com/en-us/research/publication/ensemble-deep-learning-for-speech-recognition/.

[18] Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. Learning hierarchical features for scene labeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1915–1929, 2012.

[19] Manuel Gil-Martín, Javier López-Iniesta, Fernando Fernández-Martínez, and Rubén San-Segundo. Reducing the impact of sensor orientation variability in human activity recognition using a consistent reference system. *Sensors*, 23(13):5845, 2023.

[20] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[21] Saurabh Gupta. Deep learning based human activity recognition (har) using wearable sensor data. *International Journal of Information Management Data Insights*, 1(2):100046, 2021.

[22] Saurabh Gupta. Deep learning based human activity recognition (har) using wearable sensor data. *International Journal of Information Management Data Insights*, 1(2):100046, 2021. ISSN 2667-0968. https://doi.org/https://doi.org/10.1016/j.jjimei.2021.100046. URL https://www.sciencedirect.com/science/article/pii/S2667096821000392.

[23] Sojeong Ha and Seungjin Choi. Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors. In *2016 International Joint Conference on Neural Networks (IJCNN)*, pages 381–388, 2016. https://doi.org/10.1109/IJCNN.2016.7727224.

[24] Sahand Hajifar, Saeb Ragani Lamooki, Lora A. Cavuoto, Fadel M. Megahed, and Hongyue Sun. Investigation of heterogeneity sources for occupational task recognition via transfer learning. *Sensors*, 21(19), 2021. ISSN 1424-8220. https://doi.org/10.3390/s21196677. URL https://www.mdpi.com/1424-8220/21/19/6677.

[25] Nils Y Hammerla, Shane Halloran, and Thomas Plötz. Deep, convolutional, and recurrent models for human activity recognition using wearables. *arXiv preprint arXiv:1604.08880*, 2016.

[26] Harish Haresamudram, Irfan Essa, and Thomas Plötz. Assessing the state of self-supervised human activity recognition using wearables. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 6(3), sep 2022. https://doi.org/10.1145/3550299. URL https://doi.org/10.1145/3550299.

[27] Sepp Hochreiter. The vanishing gradient problem during learning recurrent neural nets and problem solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 6(02):107–116, 1998.

[28] Zhiqing Hong, Zelong Li, Shuxin Zhong, Wenjun Lyu, Haotian Wang, Yi Ding, Tian He, and Desheng Zhang. Crosshar: Generalizing cross-dataset human activity recognition via hierarchical self-supervised pretraining. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 8(2), may 2024. https://doi.org/10.1145/3659597. URL https://doi.org/10.1145/3659597.

[29] Masaya Inoue, Sozo Inoue, and Takeshi Nishida. Deep recurrent neural network for mobile human activity recognition with high throughput. *Artificial Life and Robotics*, 23:173–185, 2018.

[30] Robert A Jacobs. Increased rates of convergence through learning rate adaptation. *Neural networks*, 1(4):295–307, 1988.

[31] Majid Janidarmian, Atena Roshan Fekr, Katarzyna Radecka, and Zeljko Zilic. A comprehensive analysis on wearable acceleration sensors in human activity recognition. *Sensors*, 17(3): 529, 2017.

[32] Ali Olow Jimale and Mohd Halim Mohd Noor. Subject variability in sensor-based activity recognition. *Journal of Ambient Intelligence and Humanized Computing*, 14(4):3261–3274, 2023.

[33] Manjunath Jogin, Mohana, M S Madhulika, G D Divya, R K Meghana, and S Apoorva. Feature extraction using convolution neural networks (cnn) and deep learning. In *2018 3rd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, pages 2319–2323, 2018. https://doi.org/10.1109/RTEICT42901.2018.9012507.

[34] David S. Johnson and Sascha Grollmisch. Techniques improving the robustness of deep learning models for industrial sound analysis. In *2020 28th European Signal Processing Conference (EUSIPCO)*, pages 81–85, 2021. https://doi.org/10.23919/Eusipco47968.2020.9287327.

[35] Jenario Johnson, Eric Williams, Micheal Swindon, Kendon Ricketts, Behzad Mottahed, Sherif Rashad, and Ryan Integlia. A wearable mobile exergaming system for activity recognition and relaxation awareness. In *2019 IEEE International Systems Conference (SysCon)*, pages 1–5. IEEE, 2019.

[36] Andrej Karpathy, Justin Johnson, and Li Fei-Fei. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078*, 2015.

[37] Azhar Ali Khaked, Nobuyuki Oishi, Daniel Roggen, and Paula Lago. Investigating the effect of orientation variability in deep learning-based human activity recognition. In *Adjunct*

Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing, UbiComp/ISWC '23 Adjunct, page 480–485, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400702006. https://doi.org/10.1145/3594739.3610742. URL https://doi-org.lib-ezproxy.concordia.ca/10.1145/3594739.3610742.

[38] Azhar Ali Khaked, Nobuyuki Oishi, Daniel Roggen, and Paula Lago. In shift and in variance: Assessing the robustness of har deep learning models against variability. *MDPI Sensors*, 2024.

[39] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[40] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Sara Beery, Jure Leskovec, Anshul Kundaje, Emma Pierson, Sergey Levine, Chelsea Finn, and Percy Liang. WILDS: A benchmark of in-the-wild distribution shifts. *CoRR*, abs/2012.07421, 2020. URL https://arxiv.org/abs/2012.07421.

[41] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[42] Prabhat Kumar and S Suresh. Deep-har: an ensemble deep learning model for recognizing the simple, complex, and heterogeneous human activities. *Multimedia Tools and Applications*, 82(20):30435–30462, 2023.

[43] TL Li, Antoni B Chan, and AH Chun. Automatic musical pattern feature extraction using convolutional neural network. *Genre*, 10(2010):1x1, 2010.

[44] Rex Liu, Albara Ah Ramli, Huanle Zhang, Erik Henricson, and Xin Liu. An overview of human activity recognition using wearable sensors: Healthcare and artificial intelligence. In *International Conference on Internet of Things*, pages 1–14. Springer, 2021.

[45] Shengzhong Liu, Shuochao Yao, Jinyang Li, Dongxin Liu, Tianshi Wang, Huajie Shao, and Tarek Abdelzaher. Giobalfusion: A global attentional deep learning framework for multisensor information fusion. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1):1–27, 2020.

[46] Sakorn Mekruksavanich and Anuchit Jitpattanakul. Multimodal wearable sensing for sport-related activity recognition using deep learning networks. *Journal of Advances in Information Technology*, 13, 2022.

[47] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Interspeech*, volume 2, pages 1045–1048. Makuhari, 2010.

[48] Chulhong Min, Akhil Mathur, Alessandro Montanari, and Fahim Kawsar. An early characterisation of wearing variability on motion signals for wearables. In *Proceedings of the 2019 ACM International Symposium on Wearable Computers*, ISWC '19, page 166–168, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450368704. https://doi.org/10.1145/3341163.3347716. URL https://doi.org/10.1145/3341163.3347716.

[49] Jose G Moreno-Torres, Troy Raeder, Rocío Alaiz-Rodríguez, Nitesh V Chawla, and Francisco Herrera. A unifying view on dataset shift in classification. *Pattern recognition*, 45(1): 521–530, 2012.

[50] Hassan Najadat, Maad Ebrahim, Mohammad Alsmirat, Obadah Shatnawi, Mohammed Nour Al-Rashdan, and Ahmad Al-Aiad. *Investigating the Classification of Human Recognition on Heterogeneous Devices Using Recurrent Neural Networks*, pages 67–80. Springer International Publishing, Cham, 2021. ISBN 978-3-030-51070-1. https://doi.org/10.1007/978-3-030-51070-1_4. URL https://doi.org/10.1007/978-3-030-51070-1_4.

[51] Maryam M Najafabadi, Flavio Villanustre, Taghi M Khoshgoftaar, Naeem Seliya, Randall Wald, and Edin Muharemagic. Deep learning applications and challenges in big data analytics. *Journal of big data*, 2:1–21, 2015.

[52] Francisco Javier Ordóñez and Daniel Roggen. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1), 2016. ISSN 1424-8220. https://doi.org/10.3390/s16010115. URL https://www.mdpi.com/1424-8220/16/1/115.

[53] Liwen Ouyang and Aaron Key. Maximum mean discrepancy for generalization in the presence of distribution and missingness shift. *CoRR*, abs/2111.10344, 2021. URL https://arxiv.org/abs/2111.10344.

[54] Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Comput. Surv.*, 55(3), February 2022. ISSN 0360-0300. https://doi.org/10.1145/3494672. URL https://doi.org/10.1145/3494672.

[55] Lionel Pigou, Aäron Van Den Oord, Sander Dieleman, Mieke Van Herreweghe, and Joni Dambre. Beyond temporal pooling: Recurrence and temporal convolutions for gesture recognition in video. *International Journal of Computer Vision*, 126:430–439, 2018.

[56] Samira Pouyanfar, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, and S. S. Iyengar. A survey on deep learning: Algorithms, techniques, and applications. *ACM Comput. Surv.*, 51(5), sep 2018. ISSN 0360-0300. https://doi.org/10.1145/3234150. URL https://doi.org/10.1145/3234150.

[57] Lumpapun Punchoojit and Nuttanont Hongwarittorrn. A comparative study on sensor displacement effect on realistic sensor displacement benchmark dataset. In *Recent Advances in Information and Communication Technology 2015: Proceedings of the 11th International Conference on Computing and Information Technology (IC2IT)*, pages 97–106. Springer, 2015.

[58] Elangovan Ramanujam, Thinagaran Perumal, and S Padmavathi. Human activity recognition with smartphone and wearable sensors using deep learning techniques: A review. *IEEE Sensors Journal*, 21(12):13029–13040, 2021.

[59] Attila Reiss. PAMAP2 Physical Activity Monitoring. UCI Machine Learning Repository, 2012. DOI: https://doi.org/10.24432/C5NW2H.

[60] Daniel Roggen, Alberto Calatroni, Long-Van Nguyen-Dinh, Ricardo Chavarriaga, and Hesam Sagha. OPPORTUNITY Activity Recognition. UCI Machine Learning Repository, 2010. DOI: https://doi.org/10.24432/C5M027.

[61] Daniel Roggen, Kilian Förster, Alberto Calatroni, Andreas Bulling, and Gerhard Tröster. On the issue of variability in labels and sensor configurations in activity recognition systems. In *Proc. "How to do good activity recognition research? Experimental methodologies, evaluation metrics, and reproducibility issues" (Pervasive)*, pages 1–4, 2010.

[62] Daniel Roggen, Meir Plotnik, and Jeff Hausdorff. Daphnet Freezing of Gait. UCI Machine Learning Repository, 2010. DOI: https://doi.org/10.24432/C56K78.

[63] Daniel Roggen, Arash Pouryazdan, and Mathias Ciliberto. Poster: Bluesense - designing an extensible platform for wearable motion sensing, sensor research and iot applications. In *Proceedings of the 2018 International Conference on Embedded Wireless Systems and Networks*, EWSN '18, page 177–178, USA, 2018. Junction Publishing. ISBN 9780994988621.

[64] Charissa Ann Ronao and Sung-Bae Cho. Human activity recognition with smartphone sensors using deep learning neural networks. *Expert systems with applications*, 59:235–244, 2016.

[65] Reuven Y Rubinstein and Dirk P Kroese. *The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation, and machine learning*, volume 133. Springer, 2004.

[66] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[67] Tara N Sainath, Oriol Vinyals, Andrew Senior, and Haşim Sak. Convolutional, long short-term memory, fully connected deep neural networks. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4580–4584. Ieee, 2015.

[68] Allan Stisen, Henrik Blunck, Sourav Bhattacharya, Thor Siiger Prentow, Mikkel Baun Kjærgaard, Anind Dey, Tobias Sonne, and Mads Møller Jensen. Smart devices are different: Assessing and mitigatingmobile sensing heterogeneities for activity recognition. In *Proceedings of the 13th ACM Conference on Embedded Networked Sensor Systems*, SenSys '15, page 127–140, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450336314. https://doi.org/10.1145/2809695.2809718. URL https://doi-org.lib-ezproxy.concordia.ca/10.1145/2809695.2809718.

[69] Sungho Suh, Vitor Fortes Rey, and Paul Lukowicz. Wearable sensor-based human activity recognition for worker safety in manufacturing line. In *Artificial Intelligence in Manufacturing: Enabling Intelligent, Flexible and Cost-Effective Production Through AI*, pages 303–317. Springer Nature Switzerland Cham, 2023.

[70] Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. Lstm neural networks for language modeling. In *Interspeech*, volume 2012, pages 194–197, 2012.

[71] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18583–18599. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/d8330f857a17c53d217014ee776bfd50-Paper.pdf.

[72] Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. Label Studio: Data labeling software, 2020-2022. URL https://github.com/heartexlabs/label-studio. Open source software available from https://github.com/heartexlabs/label-studio.

[73] Jonathan J Tompson, Arjun Jain, Yann LeCun, and Christoph Bregler. Joint training of a convolutional network and a graphical model for human pose estimation. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Cur-

ran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/e744f91c29ec99f0e662c9177946c627-Paper.pdf.

[74] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[75] LuKun Wang and RuYue Liu. Human activity recognition based on wearable sensor using hierarchical deep lstm networks. *Circuits, Systems, and Signal Processing*, 39(2):837–856, 2020.

[76] Gary Weiss. WISDM Smartphone and Smartwatch Activity and Biometrics Dataset . UCI Machine Learning Repository, 2019. DOI: https://doi.org/10.24432/C5HK59.

[77] Olivia Wiles, Sven Gowal, Florian Stimberg, Sylvestre Alvise-Rebuffi, Ira Ktena, Krishnamurthy Dvijotham, and Taylan Cemgil. A fine-grained analysis on distribution shift. *arXiv preprint arXiv:2110.11328*, 2021.

[78] Kent Wu, Suzy He, Geoff Fernie, and Atena Roshan Fekr. Deep neural network for slip detection on ice surface. *Sensors*, 20(23):6883, 2020.

[79] Miguel Xochicale, Chris Baber, and Mourad Oussalah. Understanding movement variability of simplistic gestures using an inertial sensor. In *Proceedings of the 5th ACM International Symposium on Pervasive Displays*, PerDis '16, page 239–240, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450343664. https://doi.org/10.1145/2914920.2940337. URL https://doi-org.lib-ezproxy.concordia.ca/10.1145/2914920.2940337.

[80] Jianbo Yang, Minh Nhut Nguyen, Phyo Phyo San, Xiaoli Li, and Shonali Krishnaswamy. Deep convolutional neural networks on multichannel time series for human activity recognition. In *Ijcai*, volume 15, pages 3995–4001. Buenos Aires, Argentina, 2015.

[81] Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. A review of recurrent neural networks: Lstm cells and network architectures. *Neural computation*, 31(7):1235–1270, 2019.

[82] Aras Yurtman and Billur Barshan. Activity recognition invariant to sensor orientation with wearable motion sensors. *Sensors*, 17(8):1838, 2017.

[83] Aras Yurtman and Billur Barshan. Activity recognition invariant to sensor orientation with wearable motion sensors. *Sensors*, 17(8):1838, 2017.

[84] Piero Zappi, Clemens Lombriser, Thomas Stiefmeier, Elisabetta Farella, Daniel Roggen, Luca Benini, and Gerhard Tr öster. Activity recognition from on-body sensors: accuracy-power trade-off by dynamic sensor selection. In *Wireless Sensor Networks: 5th European Conference, EWSN 2008, Bologna, Italy, January 30-February 1, 2008. Proceedings*, pages 17–33. Springer, 2008.

[85] Ming Zeng, Le T Nguyen, Bo Yu, Ole J Mengshoel, Jiang Zhu, Pang Wu, and Joy Zhang. Convolutional neural networks for human activity recognition using mobile sensors. In *6th international conference on mobile computing, applications and services*, pages 197–205. IEEE, 2014.

[86] Shibo Zhang, Yaxuan Li, Shen Zhang, Farzad Shahabi, Stephen Xia, Yu Deng, and Nabil Alshurafa. Deep learning in human activity recognition with wearable sensors: A review on advances. *Sensors*, 22(4):1476, 2022.

[87] Yexu Zhou, Haibin Zhao, Yiran Huang, Till Riedel, Michael Hefenbrock, and Michael Beigl. Tinyhar: A lightweight deep learning model designed for human activity recognition. In *Proceedings of the 2022 ACM International Symposium on Wearable Computers*, ISWC '22, page 89–93, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450394246. https://doi.org/10.1145/3544794.3558467. URL https://doi.org/10.1145/3544794.3558467.

[88] Qiuyu Zhu and Zhengyong Wang. An image clustering auto-encoder based on predefined evenly-distributed class centroids and mmd distance. *Neural Processing Letters*, 51(2):1973–1988, 2020.