# Introduction

## Wearable sensors and HAR

Recent widespread use of wearable smart devices have led them to become essential in everyday life to many. These devices enable the collection of a vast amount of data through efficient sensors, contributing to the growth of research fields utilizing these signals. And with the similar rise and applications of machine learning models, human activity recognition (HAR) has become an increasingly popular research area. A standard commercial smartwatch, for example, can gather a plethora of human data that is essential for accurate activity recognition. The growth of human activity recognition can be attributed to the numerous known benefits physical activity has on human health. Given that physical inactivity is the fourth leading cause of global morality, and that physical activity is known to be a significant prevention to at least 35 chronic diseases, monitoring and understanding individual physical activity becomes increasingly important. As a result, accurate classification and analysis of physical activity can contribute to a deeper understanding of an individual's activity patterns and the direct impact it has on their health.

## Relevant current and past research

HAR as a field of research has evolved significantly in recent years, with new techniques and methodologies continually emerging. Effective HAR models tend to focus on key areas like the quality of the data and hardware itself, the process of feature engineering, and the selection of an appropriate model. Previous research has demonstrated that utilizing multiple biosignals in a model often yields higher, or at least more consistent, accuracy compared to single signal models, as shown in Antonio A. Aguileta et al. (2019). This is further demonstrated by Ruiting Jia et al. (2013) and Justin Gilmore (2024) where both studies use multiple biosignals, electrocardiogram (ECG) and accelerometer (ACC) data in the former and ACC, blood volume pulse (BVP), and electrodermal activity (EDA) data in the latter, and implement a signal fusion technique to significantly improve model accuracy. Regarding feature engineering, many studies employing a combination of time domain and frequency domain features have been proven successful, such as in Afzali Arani et al. (2021), where a Random Forest model was able to achieve an AOC and F-1 score of 99.91% and 97.19%, respectively. To effectively implement frequency domain analysis, a specific windowing and segmentation method must be applied to the time series data. In particular, Antonios Papaleonidas et al. (2021) applied a complex windowing technique to maximize data usage. Model selection and optimization has also been a topic of heavy focus. In the same study, weighted and fine k-NNs were some of the algorithms that achieved the best performance. Ferhat Attal et al. (2015) had similarly concluded that the k-NN classifier provides the best performance compared to a few other supervised learning models tested. According to Angelica Poli et al. (2021), model success is particularly dependent on the type of measure used, where Support Vector Machines effectively predicted the relative physical activity intensities on physiological and acceleration data while Bagged Trees performed best when only Electrodermal Activity data were used.

## Our Contribution

While previous research has been successful in utilizing biosignals to predict activity with extremely high accuracy, these studies often rely on dataset collected in controlled lab settings and use a traditional train-test data split method for cross validation. The motivation for this research is to develop a better generalizable and applicable HAR system. By focusing primarily on cross individual cases and combining and testing various techniques from previous HAR systems, our goal is to create a model that can accurately recognize an individual's activity without relying on their previous data. Predicting activity across individuals enables it to be applied to new data where labeled classifications are not available and allows it to be applicable to real-world scenarios. When faced with the challenge of validating our lab collected model against real-world data, we provide an in-depth explanation and potential differences between the two data types.

# Dataset Description

The quality of the data used for HAR models is very important to how the model is able to distinguish between different activities. To ensure clear, distinct periods of activity with no overlap or ambiguous classifications, we chose the following datasets.

## PPG dataset

We utilized the PPG-DaLiA dataset, or the PPG dataset for motion compensation and heart rate estimation in Daily Life Activities, published by Attila Reiss et al. in (2019. The dataset includes 15 individuals, each equipped with a chest-worn RespiBAN monitor and wrist-worn Empatica E4 wearable device. These devices are able to capture a variety of biosignals such as photoplethysmography (PPG), three-axis acceleration, electrodermal activity, and body temperature. For the purpose of simplicity and generalizability of our model to possible other datasets, we focused exclusively on the signals collected by the Empatica E4. This decision is based on the reliability of the Empatica E4 in long term data recordings as well as its widespread use, which increases the likelihood other experiments may contain the same or similar biosignals, therefore facilitating a broader applicability of our model. While the PPG-DaLiA dataset was initially collected for improving heart rate estimations using PPG signals, our study leverages the specific periods of labeled activity that were performed to closely mimic real-world conditions. These activities include sitting, ascending and descending stairs, playing table soccer, driving a car, walking, working on a computer, and having lunch. Additionally, the dataset includes transient periods located in between activities in which the individual was either resting or preparing for the next activity.

## Optional: MOVE dataset

Similarly, the ScientISST MOVE dataset, collected by Areias Saraiva et al. (2024), includes 17 individuals, each also equipped with the Empatica E4 device along with two other sensors worn on the chest and forearm. The participants performed natural activities that are said to be executed on a daily basis by humans, which includes lifting a chair, a greeting gesture, gesticulating while talking, jumping, walking, and running. Like mentioned before, the use of the same Empatica E4 device in PPG-DaLiA and ScientISST MOVE allows for the consistent biosignal collection needed for possible cross-dataset application, crucial in our study on developing a general human activity recognition model that can adapt to a wide range of subjects and conditions.

# Feature Engineering

In this section, we outline the data preprocessing step undertaken before introducing the model. We start by detailing the data cleaning process applied to the various used biosignals in order to ensure analyzability. Then we describe the window segmentation technique we employed to optimize data utilization. Lastly, we explain how we converted the time series data into the frequency domain to enhance feature extraction.

As mentioned in the introduction, according to Afzali Arani et al., combining ACC, ECG, and PPG data enhances the model's prediction accuracy, which is why we have chosen to utilize these features. The Empatica E4 collects these signals at different frequencies, ACC in 32 Hz, PPG in 64 Hz, and ECG or heart rate (HR) in 1 Hz. While other HAR methods may impute or downsample these frequencies for consistency, our feature engineering approach allowed us to not interfere with the data in this manner. In addition, the PPG dataset does not include any missing data and eliminates the need to address this issue.

## Windowing and Complex Windowing

To effectively implement frequency domain analysis, a specific windowing and segmentation method must be applied to the time series data.

Common to many other HAR projects, we segment the data, as analyzing time series data in windows is more meaningful that each datapoint individually. By using a sliding window, we can section potions of the data at a time and use these windows as new data instances. In the time domain, we extract features from each window, creating a new dataset that can be used to train our model. This approach allows us to leverage the segmented data effectively, ensuring that our model is trained on meaningful and representative data.

According to Afzali Arani et al. (2021), the window length and offsets are very important to collect meaningful and accurate instances of the activity and must be carefully selected depending on the activity and its duration. Windows must be large enough to capture periodic activity signals, especially with our method of feature engineering in the frequency domain, however large windows may result in delays in real-time signal processing or include multiple activity classifications in a single window. While various different window durations and offsets have been implemented in the past, anywhere between 0.08 to 30 seconds, our observations with window lengths between 3 to 8 seconds and offsets ranging from 1 to 5 seconds resulted in insignificant improvement in our HAR model's performance. Consequently, we decided to use 7-second windows with no overlapping windows (7 second offset) following the window parameters recommended by them.

When applying this segmentation method, some windows may contain data during transitional periods, resulting in multiple classified activities appearing within a single window. Inspired by the complex windowing technique implemented by Antonios Papaleonidas et al. (2021), we implemented a proportion threshold to ensure that the majority of the data within each window corresponds to a single activity. Although we set the threshold at 1%, this value can be adjusted as needed.

## Fourier Transformation

Within these windows, we further improve our features by implementing a Fourier Transform, a technique widely used in many HAR models to convert time series data into the frequency domain. Our aim in this transformation is to collect periodic signals from the data, ones that are distinct and unique to specific activities.

The Fourier Transform decomposes wavelengths in the time domain into a sum of sinusoids of different frequencies, allowing us to analyze the frequency components of the data. The key advantage this technique has is the ability to decompose any waveform, meaning all signals, including the biosignals we are working with, can be converted into the frequency domain. The relationship between time and frequency domains is states as

Where X(f) is the function of frequency and the Fourier Transform of x(t) in the time domain.

As an approximation of the continuous Fourier Transformation, and one that is more suitable with our data, we use the discrete case of the Fourier Transform described as follows:

,

Where sequence Xk is the discrete Fourier Transform of sequence xt of N terms. Even further, we use the Fast Fourier Transform (FFT) algorithm for a further accurate computation which computes several smaller DFTs instead of one large one. The complexity decreases significantly when DFT is partitioned, as the number of operations required to directly compute an N-point DFT is proportional to N2. This makes FFT a powerful tool, especially when N is highly composite.

Adopting the feature extraction methods outlined in Afzali Arani et al., we extract seven statistical features from the time domain in each window. These features are listed below.

* mean
* minimum
* maximum
* median
* standard deviation
* zero-crossing rate
* mean-crossing rate

These features are commonly used in many HAR models, but it is worth detailing that the zero-crossing rate and mean-crossing rate are the number of times the signal crosses zero or the mean, respectively.

With our new sequence after the Fourier Transform, we are able to collect seven more features in the frequency domain,

* mean
* minimum
* maximum (for HR and PPG) / second maximum (for ACC)
* DC component
* median
* dominant frequency
* mean-crossing rate

For further clarification, the first five features collect the amplitude value in the frequency domain. The dominant frequency, in contrast, is the frequency measure the corresponds with the maximum value (the second maximum value for ACC)

Also note the use of the maximum for HR and PPG data while the second maximum is used instead for ACC data. The DC component, which can be understood as the amplitude value at zero, is heavily impacted by how the Empatica E4 device measures acceleration. Similar to many accelerometers, there is a continuous effect of gravity the Empatica E4 measures, which significantly impacts the amplitude value at zero, or the average value of signal in time-domain, resulting in the DC component appearing as the maximum value. Therefore, the second maximum value is collected for better representation of the window and to remove redundancy in the features.

# Results and Limitations

## Initial Threshold Filtering

Independent from the windowing technique and feature engineering methods previously discussed, our initial analysis of the PPG-DaLia dataset focused on addressing the extreme variability present in the raw ACC data alone. Accelerometer data is highly variable, particularly when measuring dynamic human movements in their raw form. Combined with the fact that these human movements are unclassified in the PPG-DaLia dataset, it becomes increasingly difficult to derive insights without proper preprocessing or noise reducing techniques.

We begin by using the Fourier Transformation to filter out unnecessary signals from the data, then using the denoised results to train a KNN model. We opt to employ a dynamic filtering technique, selecting the top 0.01% of signals by amplitude for each dimension in the accelerometer. This approach retains approximately 17 signals, all with amplitudes greater than 51000. Training a K-Nearest Neighbour model with these signals instead achieves an accuracy greater than 99.9% when validating on a split test set.

However, validating the model using Leave-One-Individual-Out Cross Validation (LOOCV) , a method that is critical when evaluating subject independent models, results in a significant decline of model accuracy. This reduction is likely due to the aggressive denoising process which resulted in a significant lack of variance, one essential to balancing bias and preventing overfitting. Nevertheless, the reduction of noise within each individual’s data was expected to reduce subject-specific variability, potentially improving generalizability. While our method may have been overly aggressive, it still effectively illustrates that developing a generalized model for human activity recognition may be more complex than initially assumed. The way one individual performs a variety of tasks may be entirely different from another, making it difficult to train a model in such a manner. The inherent variability in natural human movements suggests that relying solely on ACC data may not be sufficient for accurately predicting activity.

## Simple Windowed Models

Continuing with the feature engineering methods previously described, this section will focus on the results obtained with regards to the previously mentioned windowing parameters of 7 second lengths and no overlap. We implement three well known supervised learning models, a k-nearest neighbor (KNN) , a Gaussian support vector machine (GSVM), and random forest (RF) , which are similar to many other methods used by previous HAR research in order to replicate their outcomes. Like mentioned before, we focus on LOOCV as validation in order to gain insight on generalizability.

We test various combinations of activity classes for the sake of simplicity. Many HAR models focus on predicting specific activities while others rather focus on accurately classifying low mid high intensity exercise. While we can collapse many of these classes to one simple class of “activity” or “no activity”, we must be aware of the implications. Periodic signals such as walking and running may be fine to collapse into one class, but the classification of “no activity” must have more nuisance.

While a baseline measure is the simplest form of “no activity”, adding additional classifications unique to the PPG-DaLia dataset may help our model develop more insights and more impactful outcomes. As a result, we define no\_activity and transient activity as separate classes unique to the PPG\_DaLia dataset, where no\_activity is a collective of all labels classifications given by the dataset that aren’t quite defined as physically activity, which includes baseline, working, lunch, and driving. Transient activity, on the other hand, is named after the transient periods in between labeled periods of time. Considering our motivation to create a generalized model to identify physical activity in real world data, having periods of essentially random noise to our model was an added challenge. As shown in our model results, both of these classifications are tested.

Some important information about these classifications can be seen with each iteration of model results. Cycling is unique due to our data being collected by a wrist worn device, resulting in a high intensity workout with little ACC variability compared to other activities like table soccer. Alternatively, classifying table soccer poses significant challenges due to the limited presence of periodic or repetitive movements that can be learned and predicted, which can limit the effectiveness of our feature engineering using the Fourier transform.

DESCRIBE KNN

K neighterst neights bnor a classification algorith to caterogize datapoints by measuring the similarities to other dat within the set. More particalary, using distance metres, mainl;y euclidiat distances, it takes a equal vote of its k nearest neighbors and assigns the test datapoint the majority class. On our models, we run maultiple k nn models, optimaized for accuracy as well as precision, an different iteration of classes. We list them in the following  
DESCRIBE GSVM  
DESCRIBE RF

INCLUDE CHART OF ALL MODELS IN DETAIL WITH ACCURACY AND PRECISION

* No activity vs activity
* No activity + transient vs activity
* Transient activity vs activity
* Optimal for knn/gsvm/rf

Similar observations can be made even in different simulations. Stairs and walking are often confused with each other, which should make sense intuitively. Most common is a normal regression towards the “no activity class” especially when we include transiatn ictivity into that classification. Our most optimal solutions, model 4, shows good accuracy and precision but can be decided to an imbalance of class size that contributes to a abnormality high rating in each category.

## Complex Windowed Models

This section describes the results obtained when tested through multiple window lengths and offsets. Like mentioned before, these subtle changes were insignificant to the final result and optimal accuracy measure, majority of them either decreased in accuracy or were similar or slightly better. Out overall conclusions is that our 7 seconds 7offset was a simplestaparamtes that optimized for accuracy and ease.

## Generalization Problem

domain shift issue

General difficult generalization of human activity

## UMAP Feature Reduction

* Using less features and a similar model could help
* PCA found ineffective
* UMAP as a method of feature reduction while keeping classification distances and neighborhoods relevent
* Slightly increases accuracy
* Lmiatation is that is it unsupervised and may be difficult to interpret when introducing unclassified data
* Note: Supervised UMAP was another method that resulted in incredibly high accuracy, but was overfit to the trained data
* Semi-supervised UMAP also return even worse accuracy

## Optional: Applying to MOVE dataset

# Conclusion and Future Work

## Cross individual data limitations

## MOVE dataset Limitations

## Future unsupervised learning or feature extraction

Ridge regression to reduce bias and create more generaliziability

Unsupervised learning models

# Sources

<https://iris.who.int/bitstream/handle/10665/44399/9789?sequence=1>

<https://onlinelibrary.wiley.com/doi/full/10.1111/sms.12581>

<https://www.mdpi.com/1424-8220/21/21/6997#sec5dot2-sensors-21-06997> - uses acc, ppg, and ecg for improving / classification method

<https://www.tandfonline.com/doi/epdf/10.1080/24751839.2021.1987706?needAccess=true> - using a complex windowing technique for semeting

<https://www.mdpi.com/1424-8220/16/10/1569> - biomedical application of smart sensors

<https://www.mdpi.com/1424-8220/19/17/3808> - biometric fusion techniques

<https://www.mdpi.com/1424-8220/24/10/3005#B20-sensors-24-03005> - uses acc, bvp, and eda for classification / could be used for different biosignals for the classification

<https://ieeexplore.ieee.org/abstract/document/6664056> - ecg data enhances classification that uses just acc

<https://physionet.org/content/scientisst-move-biosignals/1.0.1/> - MOVE dataset

<https://ieeexplore.ieee.org/abstract/document/6365160> - HAR overview

<https://boa.unimib.it/bitstream/10281/265204/1/08995531.pdf> - idk needs to be read, possibly important for our concluding findings