Extending IMU-Based Human Activity Recognition with Cross-Subject Adaptation, Early Activity Recognition & Fatigue Estimation

by

Zarif Ashrafee
22301222
ORCID: 0009-0003-9439-3326

Notebook URL: Google Colab

A project report submitted in partial fulfillment
of the requirements for the course
CSE424: Pattern Recognition

Department of Computer Science and Engineering
BRAC University
September 2025.

Abstract

Human Activity Recognition has conventionally focused on high accuracy multi-modal fusion for UX improvements in healthcare, fitness tracking, smart home ecosystems, security monitoring & emergency response applications. An unobtrusive method of implementing Activity Recognition is through the use of IMU sensors. Since most everyday carry smart devices contain IMU sensors, validating IMU-based HAR has become quintessential. As every subject may not perform a particular activity in the exact same manner, exploring cross-subject domain adaptation techniques has been an untended area in past works. Activity forecasting via Early Activity Recognition and Fatigue Estimation through the analysis of physiological load can also improve upon the existing frameworks of HAR.

Keywords: human activity recognition (HAR), inertial measurement unit (IMU), cross-subject domain adaptation, early activity recognition, fatigue estimation

Contents

Chapter 1

Introduction

1.1 Background

Human Activity Recognition (HAR) has become more important than ever as wearable sensors continue growing as a part of everyday life. With the rapid increase of fitness trackers, smartwatches and medical monitoring devices, identifying human activities accurately is crucial for rehabilitation, eldercare, fall detection and personal fitness tracking applications. This paper aims to provide both baseline models and practical extensions that make HAR more applicable in real-world situations.

1.2 Problem Statement

Many studies show high accuracy on test datasets but real-world implementations often struggle due to the common challenges of inter-subject variability, latency, sensor robustness & limited interpretability. Models trained on a certain group of individuals may not perform well with other demographics. Also, classifying a full window requires several seconds of data, which limits the real-time response. Furthermore, real deployments can face issues such as sensor disconnections, dead batteries, or misplacement. Lastly, black-box deep learning models also offer very little in terms of high-level insights, which can reduce trust in healthcare settings.

1.3 Objective

Addressing the key research questions above, the primary objectives of this study are:

- Address subject variability using domain adaptation methods like CORAL & DANN.
- Investigate early activity recognition by predicting activities from partial sequences to reduce latency.
- Incorporate physiological signals by designing a multitask model that predicts both activity and exertion.
- Evaluate robustness to missing sensors through simulated sensor dropouts.
- Improve interpretability using SHAP & LIME to identify sensor contributions.

1.4 Scopes & Challenges

This paper does not aim to excel at high accuracy HAR as found in many multi-modal HAR studies and rather approaches the aforementioned problems of existing IMU-based solutions.
Currently, beyond the scope of this study is exploring more sophisticated Deep Neural Network-based (DNN) domain adaptation strategies, Sequence to Sequence Neural Networks for activity forecasting & transitions, exploring the effect of multi-modal fusion by comparing an IMU-only baseline and deriving actual exertion labels from participant surveys for physiological load estimation.
Subject information provided in the PAMAP2 dataset has been attached hereby for understanding the demographic spread. For cross-subject adaptation techniques, a wider demographic may be explored for generalization across the board.

Subject ID	Sex	Age (years)	Height (cm)	Weight (kg)	Resting HR (bpm)	Max HR (bpm)	Dominant hand
101	Male	27	182	83	75	193	right
102	Female	25	169	78	74	195	right
103	Male	31	187	92	68	189	right
104	Male	24	194	95	58	196	right
105	Male	26	180	73	70	194	right
106	Male	26	183	69	60	194	right
107	Male	23	173	86	60	197	right
108	Male	32	179	87	66	188	left
109	Male	31	168	65	54	189	right

Chapter 2

Literature Review

2.1 Preliminaries

Human Activity Recognition

Human Activity Recognition is the task of automatically identifying physical activities from sensor observations [6]. By analyzing available on-device sensor streams, user activity can be determined via leveraging Machine Learning techniques.

Sliding Windows

Sliding windows are used to segment sequential data such as time-series data into prefixed sizes, where the window is able to move across the data step-by-step. Since random data points of time-series data may not hold much correlation with the sensor signature of an activity, segmenting the data into sizable chunks can yield recognizable patterns for the machine to learn from.

Correlation Alignment

Correlation alignment takes the covariance matrices from the subject and the target, finds & applies a linear transformation to minimize differences and transforms the source domain to align it with the target. This minimizes subject-induced biases and prevents the machine from learning traits that may not generalize well across a wider test set [5].

Domain-Adversarial Neural Networks

Domain-Adversarial Neural Networks train a model to learn domain-invariant features by adding a domain classifier and applying a gradient reversal. During training, the model learns to perform a main task, such as classification. It also learns to trick the domain classifier, which tries to differentiate between the source and the target domains. This adversarial process pushes the model to extract features that are useful for the main classification task but not tied to a specific domain. As a result, performance on new, unseen data is improved and subject-induced bias is minimized[8].

Early Activity Recognition

Early activity recognition is the act of identifying user activity before it is completed by using prior observed data. Early recognition considers truncated windows of the original segmented dataset to extract features that align the most with the target. This reduces latency in real-time applications requiring HAR [2], as it shrinks data required to make a prediction.

Physiological Load

Physiological load refers to a user's exertion levels. Typically measured through participant surveys asking them to rate their experienced fatigue on a scale of 1 to 10, or through EEG sensors [7], gathering precise exertion levels is a challenge in the HAR space due to dataset limitations and due to on-device sensor availability, as most target devices do not incorporate EEG sensing techniques. However, heart rate (HR) can complement inertial data by reflecting a direct correlation to exertion.

2.2 Overview of the Key Findings

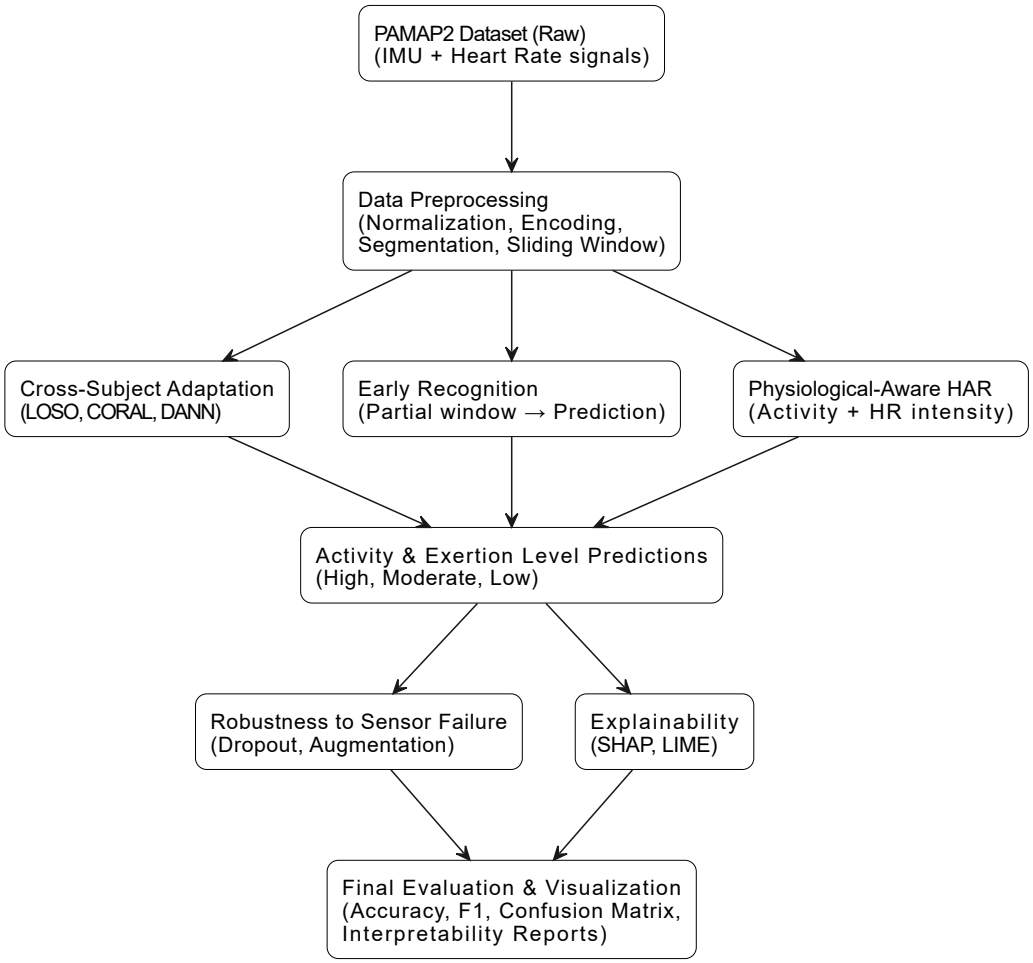
Conformer based [3] & hierarchical [4] classifications for HAR have been explored, which divide all human activity into Static & Dynamic activities. As hierarchical classifications differ from conventional classifiers, a local & global classifier approach was tested. This model has also been generalized across various IMU-based HAR datasets. IMU-encoders have also been pretrained with multi-modal self-supervision [6] in prior work, paving the way for wearable devices with multiple streams of data for activity recognition.
With the AHDT system beating state-of-the-art classifiers in accuracy, it is clearly not optimized for latency or computational efficiency, as the decision tree becomes complex with the introduction of new activities, as shown by the authors. The model also treats all instances of a detected activity as equal, ignoring motion variability or signal degradation over time, hence no fatigue or effort estimation is quantified unlike in works which work solely in identifying user fatigue [1], [7]. The classifier also labels static activities only and overlooks transitions (e.g., sit-to-stand, walk-to-run), which carry key contextual cues. Lastly, the hierarchical classifier assumes activities are independent, missing important temporal correlations between actions. Thus, it cannot anticipate what a user is likely to do next, which restricts proactive behavior in smart systems.

Chapter 3

Methodology

3.1 Methodology Overview

A flowchart of the proposed methodology:

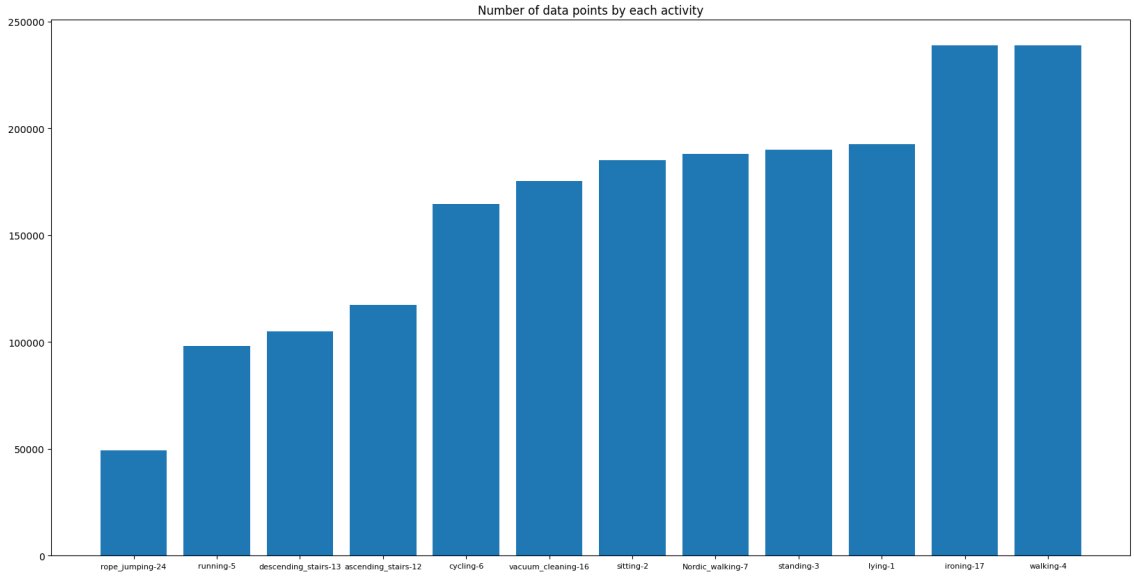


3.2 Data Collection

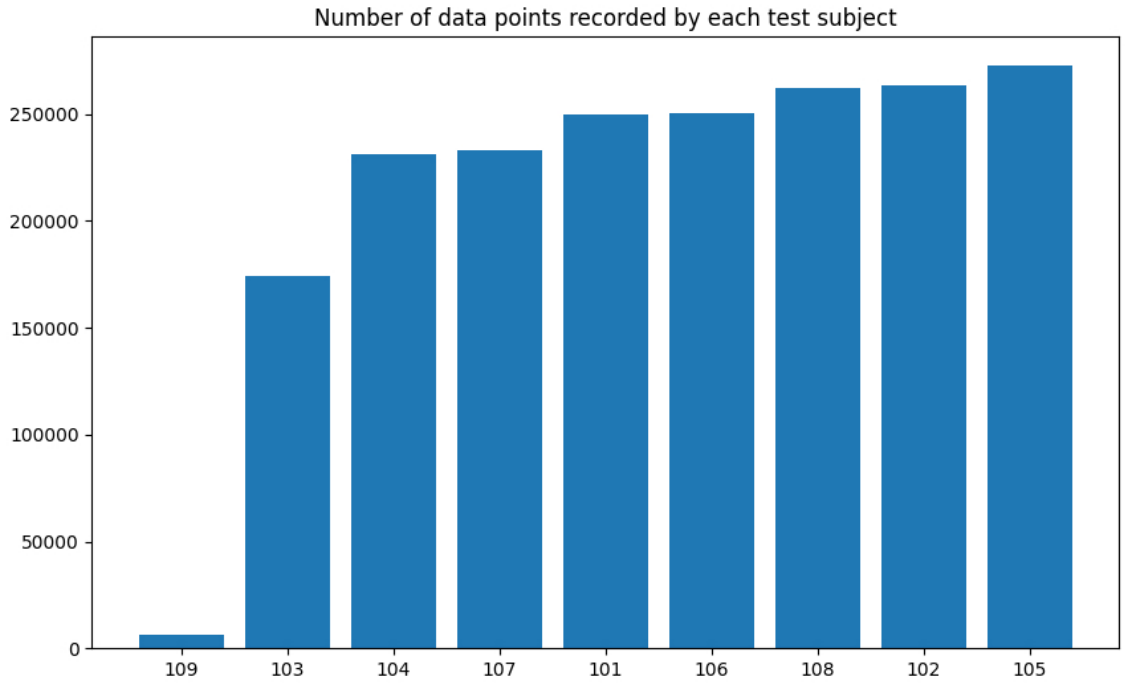
The PAMAP2 dataset contains 2,872,533 rows × 55 columns including timestamps, activity ID (25 mapped IDs), heart rates in BPM, 3 IMU sensor groups (hand, chest & ankle). The IMU sensors contain temperature readings in Celsius, 3D acceleration data in both 6g and 16g, 3D gyroscope data, 3D magnetometer data and the orientation.
As per the FAIR & CARE principles, the PAMAP2 dataset is open, reproducible and interoperable. PAMAP2 is also anonymized and subject details cannot be traced back to personal identification factors.

3.2.1 Exploratory Data Analysis

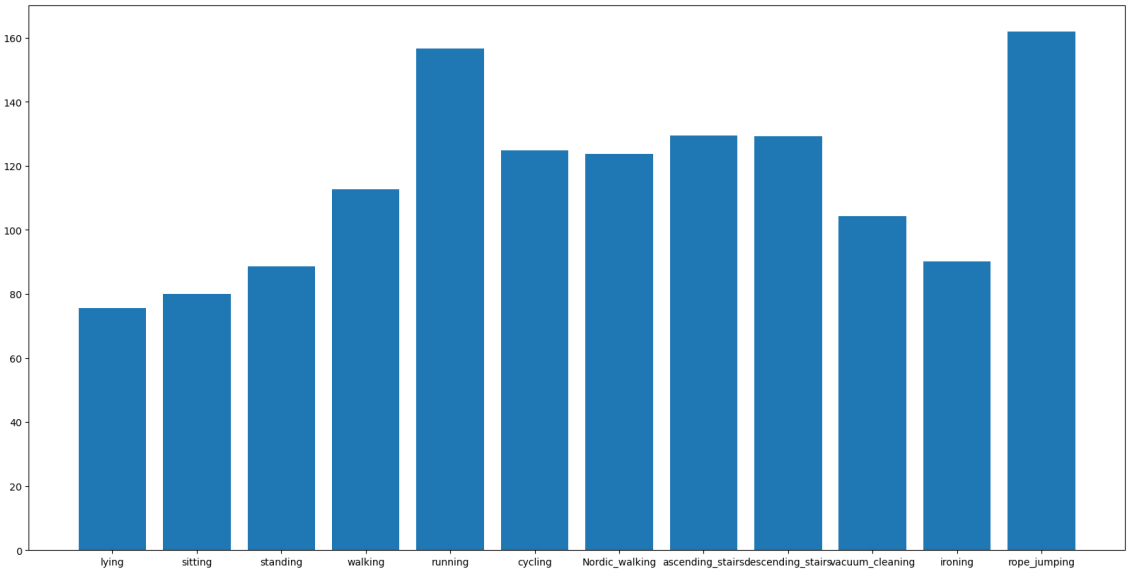
Visualizing the dataset in the time-domain, an imbalance can be observed in the target label. Not all activities have an equal number of data points; thus synthetic minority oversampling technique (SMOTE) was used to interpolate a balanced number of data points in the target.



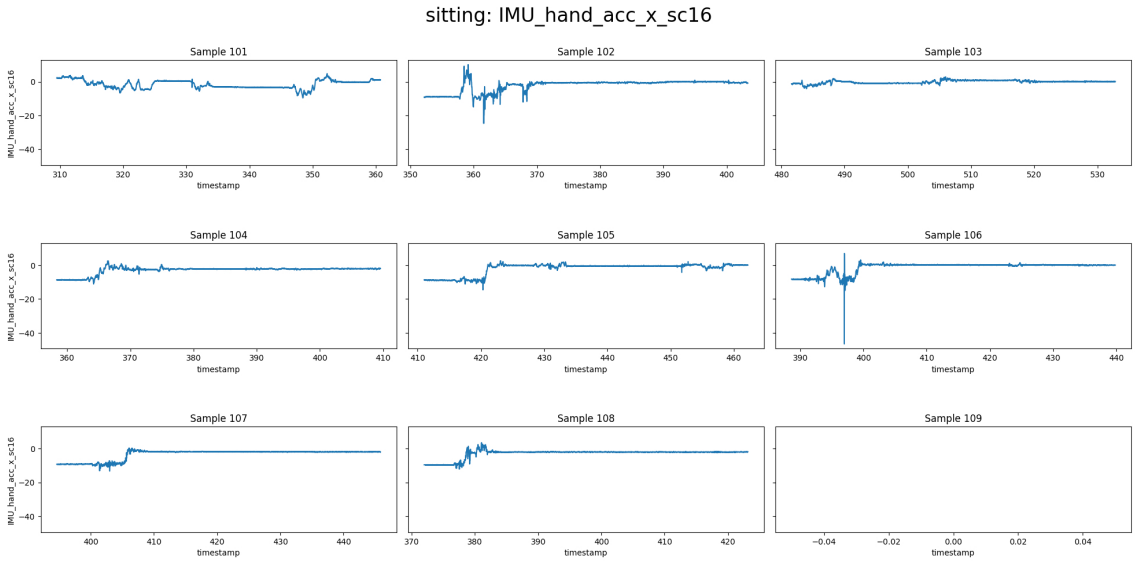
Furthermore, not all participants had an equal number of entries. Since subjects are not the target label, they do not require synthetic generation of equal data points. However, this needs to be addressed during cross-subject adaptation techniques when manually assigning particular subjects to test-train subsets.



Analyzing the mean heart rate distribution across all activities to identify high-intensity activities, rope jumping and running seem to have the highest BPM values across the board. This insight denotes a higher level of exertion when performing those activities.



Looking at one axis of raw sensor acceleration data in 16g plotted against time, signal spikes, noise, dropouts and transients can be observed. This requires addressing during the dataset pre-processing techniques.

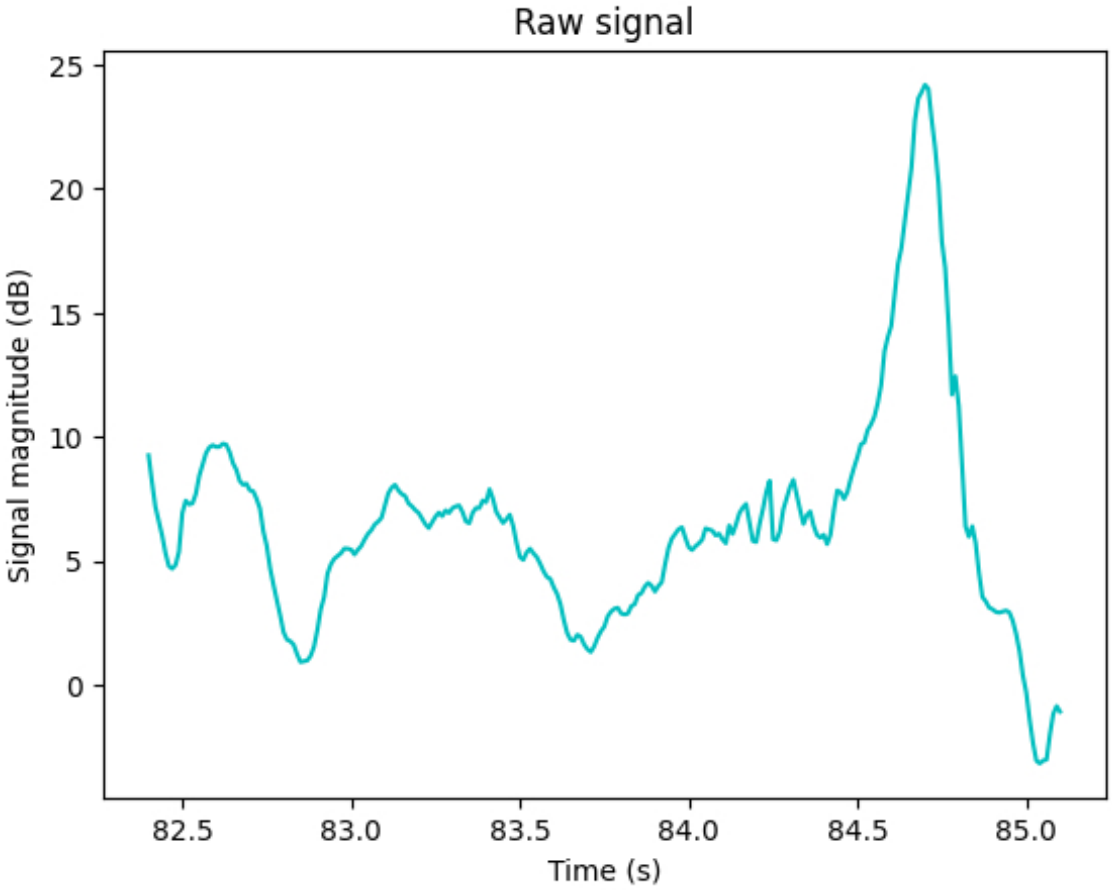


3.2.2 Data Cleaning

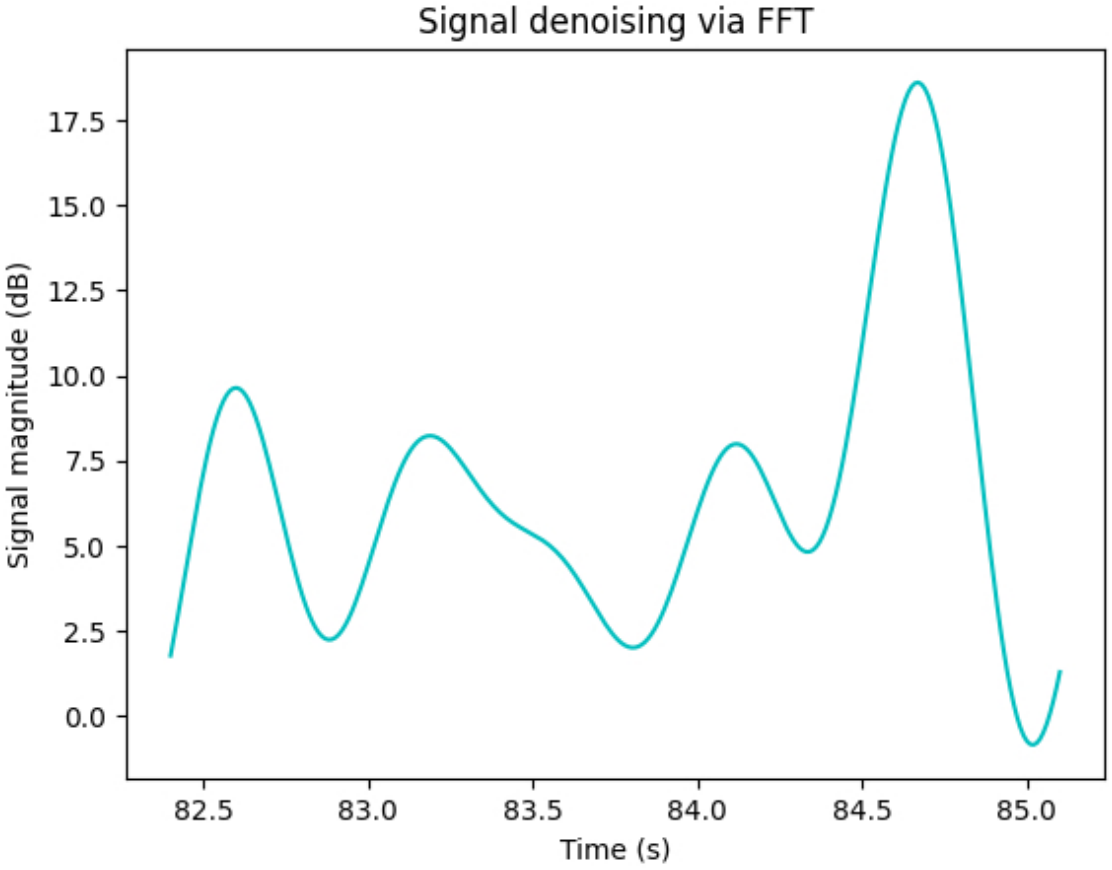
As per the dataset documentation, Activity ID '0' denotes transient activities or breaks and therefore has been dropped. Due to sensor dropouts & sampling rate mismatches (9 Hz on the Heart Rate sensor vs. 100 Hz on the IMU sensors), over 500,742 null values have been imputed (406,205 just from heart rates) via the mean values of the corresponding columns.

3.2.3 Data Transformation

Sliding window sizes of 5.12 seconds or 512 samples were chosen to segment the dataset. The selected window shifted by 1 second between consecutive segments.



Power Spectral Density (PSD) was used to filter out noisy signals from the segmented data. A low-pass filter was applied to retain only the frequencies within 100 Hz and small Fourier coefficients were zeroed out.



3.2.4 Data Reduction

The first and last 10 seconds of each labeled activity have been trimmed due to inactivity and eventual transients. This resulted in faster training times and removed empty placeholders for data irrelevant to the performed activity.

3.3 Model Specifications

3.3.1 Baseline Comparison

Baseline models were trained using Logistic Regression & Random Forest Classifier. For the Logistic Regression model, hyperparameter max iteration was set to 500, multi-class classification was enabled via 'multinomial' and the solver was set to 'saga'. For the Random Forest Classifier, n estimators were set to 200, the max depth was set to 20, class weights were set to 'balanced-subsample', a random state was set at 0 for reproducibility of the results and n jobs was set to -1 to utilize all available cores for training.
Label Encoder was used to encode all target classes. Standard Scaler was used to normalize the values of all X columns. Cross-subject validation on traditional ML models was implemented via the LOSO (Leave-One-Subject-Out) method. As Subject 108 had a balanced participation, they were chosen as the Target column for the Test set and all other subjects other than Subject 109 were set as the source for the Train set. Subject 109 had very limited participation and therefore was discarded.

3.3.2 CORAL & DANN

Since CORAL needs to be trained on a model, the baseline Logistic Regression classifier was chosen for multi-class classification. The same hyperparameter sweep was kept for comparing the findings to the baseline. Only 10 out of the 12 activities have been tested to reduce training time. This left 4205 rows of the segmented dataset.
For DANN, a gradient reversal layer makes the feature extractor 'confuse' the domain classifier. Sequential Neural Networks with Linear (64, 2) & ReLU activation functions were used for both the feature extractor & the label predictor. A batch size of 256 was selected and training was compared for 8, 10 & 20 epochs, with 10 epochs yielding optimized results.

3.3.3 Fatigue Estimation

Since the PAMAP2 dataset does not contain actual exertion levels from participant surveys, the heart rate feature was chosen as the closest correlated feature, as observed from the heart rate distribution figure accurately portraying the two high-intensity activities.
Using a Standard Scaler, all heart rate values were normalized to a range of 0 to 1. Three exertion levels - High, Moderate & Low, were considered. These would be the new proxy target labels for fatigue estimation. With the normalized values, tertile division would quantify values below 0.33 as Low exertion level, above 0.33 and below 0.67 as Moderate exertion level and anything above or equal to 0.67 as High exertion level. The target was encoded via a Label Encoder and the same baseline Logistic Regression classifier was used to predict exertion levels for coherence.

3.3.4 Early Activity Recognition

In the segmented dataset, each window consisted of 512 samples. For early activity recognition, prefix lengths from 1 second up to the full window (100, 200, 300, 512) were tested for comparison. The segments were recreated based on the new moving windows to see how much of the segment size is required to make an accurate early prediction. The same baseline Logistic Regression classifier was used for testing early predictions with the actual target activity labels.

3.3.5 Robustness under Sensor Failure

First, each sensor group (Hand, Chest & Ankle) columns were selectively zeroed out to observe the prediction accuracy drop from the baseline. Next, a more reasonable scenario of sensor dropout was augmented via a random probabilistic approach. Each group would have a 30% chance of being dropped and 3 augmented or artificial samples would be generated per dropped sample. Lastly, improvement from the original grouped dropout to the augmented scenario was measured using the baseline Logistic Regression classifier and the random seed was set to 42 for reproducibility of the results.

3.3.6 Hardware Specifications

All models were trained on a single RTX 4060 GPU with a Ryzen 7 7840HS Processor and 16 GB of LPDDR5x RAM at 6400 MHz.

Chapter 4

Results

4.1 Overview of the Comparison

Each model was tested for accuracy, macro & weighted average of precision, recall & F1 scores. Confusion matrices were plotted wherever relevant and the explainability of the models was tested via Local Interpretable Model-agnostic Explanations (LIME) & Shapley Additive Explanations (SHAP).

4.2 Analysis of the Key Findings

4.2.1 Baseline Comparison

The Logistic Regression classifier achieves an accuracy score of **82.745826%**.

	Precision	**Recall**	**F1-Score**
Macro Average	0.88	0.86	0.83
Weighted Average	0.87	0.83	0.80

Table 4.1: Baseline Logistic Regression Classifier

The Random Forest classifier achieves an accuracy score of **94.619666%**.

	Precision	**Recall**	**F1-Score**
Macro Average	0.97	0.96	0.95
Weighted Average	0.96	0.95	0.94

Table 4.2: Baseline Random Forest Classifier

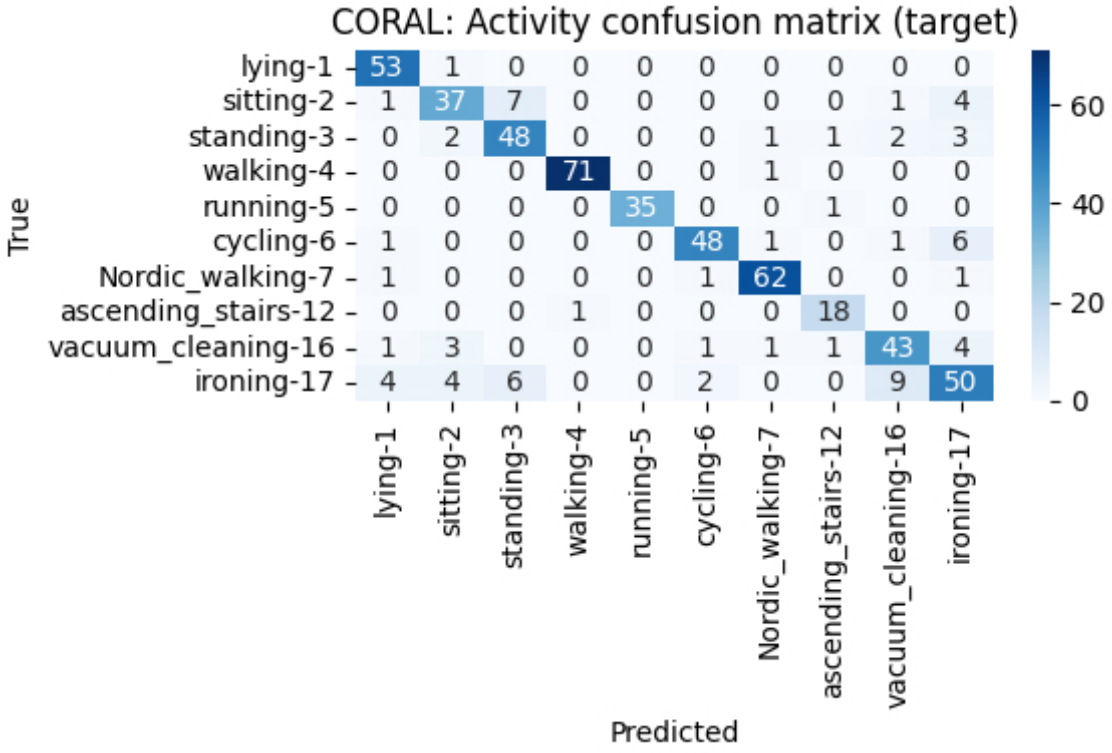
While the Random Forest classifier achieves an impressive score, the Logistic Regression classifier also does an acceptable job as a reference point for comparison. Since it is a simpler implementation, using the Logistic Regression classifier to compare scores (gains or deficits) to the baseline will be faster in terms of training time & allow for headroom when analyzing each method if there is a significant gain to be measured.

4.2.2 CORAL & DANN

Applying Correlation Alignment on a Logistic Regression classifier achieves an increased accuracy score of **86.270872%**.

	Precision	Recall	F1-Score
Macro Average	0.87	0.87	0.87
Weighted Average	0.86	0.86	0.86

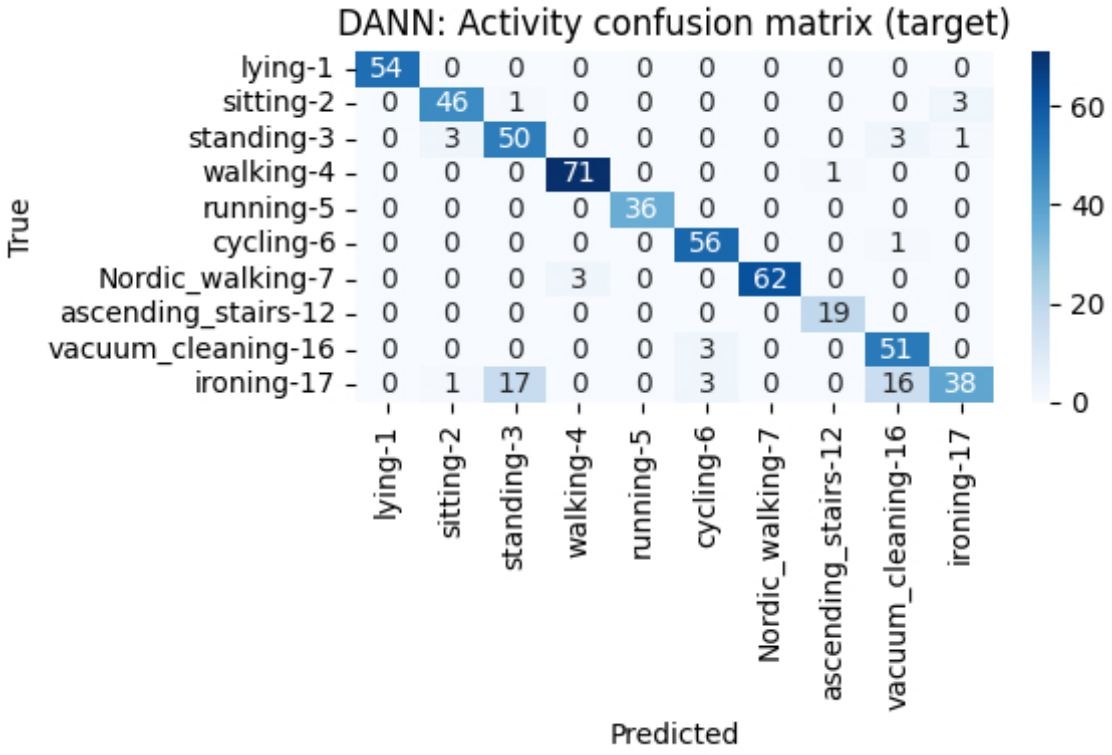
Table 4.3: Correlation Alignment



Training Domain-Adversarial Neural Networks achieved an impressive accuracy score of **89.6103896%**, signifying a substantial increase over the baseline.

	Precision	**Recall**	**F1-Score**
Macro Average	0.91	0.92	0.91
Weighted Average	0.91	0.90	0.89

Table 4.4: Domain-Adversarial Neural Network



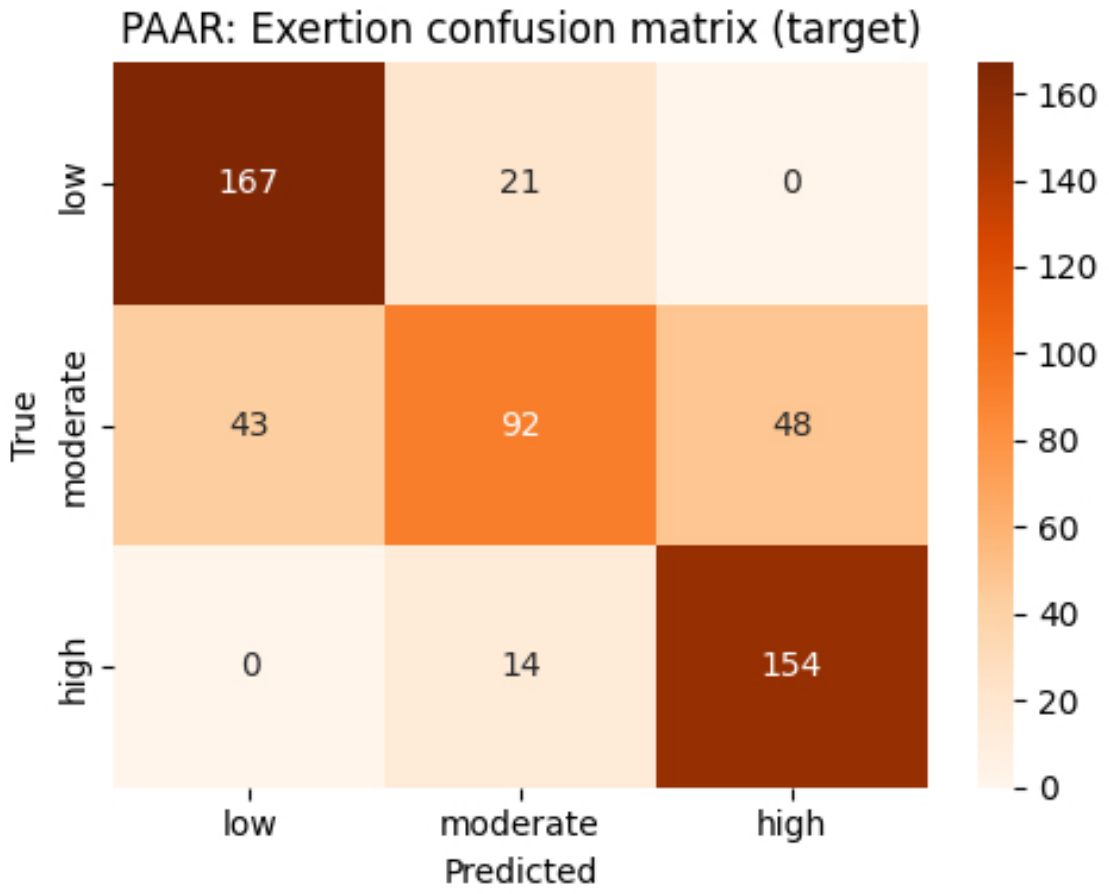
Both methods of cross-subject domain adaptation show improvements over the baseline Logistic Regression classifier. This shows the relevance of these techniques in HAR.

4.2.3 Fatigue Estimation

Training a baseline Logistic Regression classifier on the derived exertion levels achieves an accuracy score of **76.623377%**.

	Precision	Recall	F1-Score
Macro Average	0.76	0.77	0.76
Weighted Average	0.76	0.77	0.75

Table 4.5: Fatigue Estimation



While the accuracy scores are lower across the range, especially with the misclassifications of low & high exertion levels as moderate, this still adds an additional dimensionality to HAR. The ability to gauge user fatigue may contribute to UX improvements, particularly in fitness tracking & healthcare applications where the system may benefit from such predictions.

4.2.4 Early Activity Recognition

Testing different prefix lengths (from 100 to 512) on the baseline Logistic Regression classifier to find the ideal prefix length where an early prediction can be reliable. Setting the prefix length to 100, the model achieves an accuracy score of **52.133581%**.

	Precision	**Recall**	**F1-Score**
Macro Average	0.50	0.53	0.47
Weighted Average	0.47	0.52	0.44

Table 4.6: EAR: Prefix Length 100

Setting the prefix length to 200, the model achieves an accuracy score of **55.287569%**.

	Precision	**Recall**	**F1-Score**
Macro Average	0.53	0.58	0.51
Weighted Average	0.48	0.55	0.47

Table 4.7: EAR: Prefix Length 200

Increasing the prefix length to 300, the model achieves the highest accuracy score of the bunch at **55.844156%**.

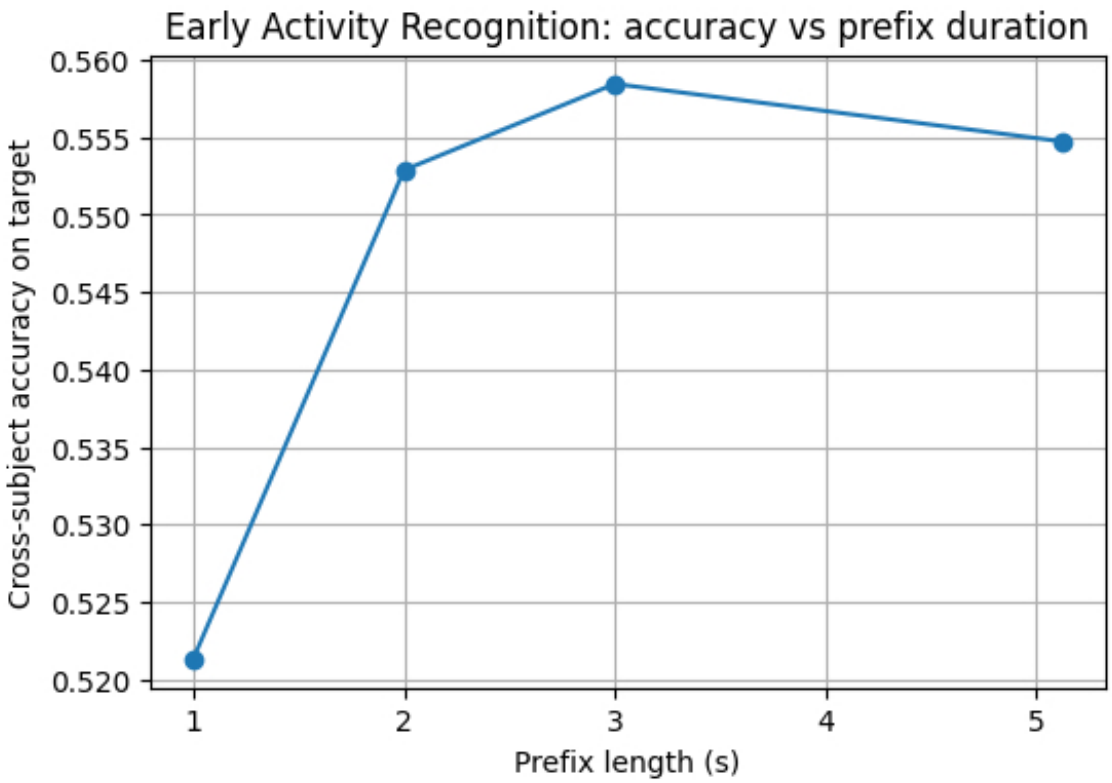
	Precision	**Recall**	**F1-Score**
Macro Average	0.55	0.59	0.51
Weighted Average	0.50	0.56	0.46

Table 4.8: EAR: Prefix Length 300

At a prefix length of 512, the model achieves an accuracy score of **55.473098%**, dropping a negligible amount from the prior prefix length.

	Precision	**Recall**	**F1-Score**
Macro Average	0.46	0.58	0.50
Weighted Average	0.39	0.55	0.45

Table 4.9: EAR: Prefix Length 512



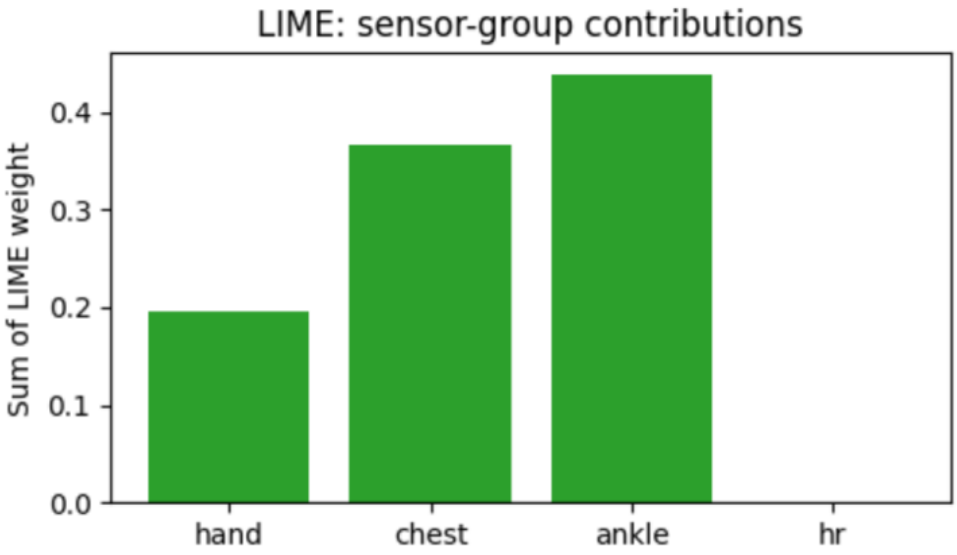
Even though a prefix length of 300 seems to yield the best results, the accuracies across the board are not very strong. Therefore, sequential models such as LSTMs must be used, as the Logistic Regression classifier can correctly predict oncoming events only about half the time, despite feeding it an entire window prior.

4.2.5 Robustness under Sensor Failure

Firstly, grouped sensor dropouts were emulated by zeroing out entire columns (Hand, Chest, Ankle) respectively. Then the described probabilistic dropout scenario was augmented and tested for each sensor group.

Sensor Group	Accuracy	Drop from Baseline	Augmented Accuracy
Hand Sensor Dropout	70.50%	9.65%	74.03%
Chest Sensor Dropout	36.36%	43.78%	53.80%
Ankle Sensor Dropout	33.02%	47.12%	91.65%

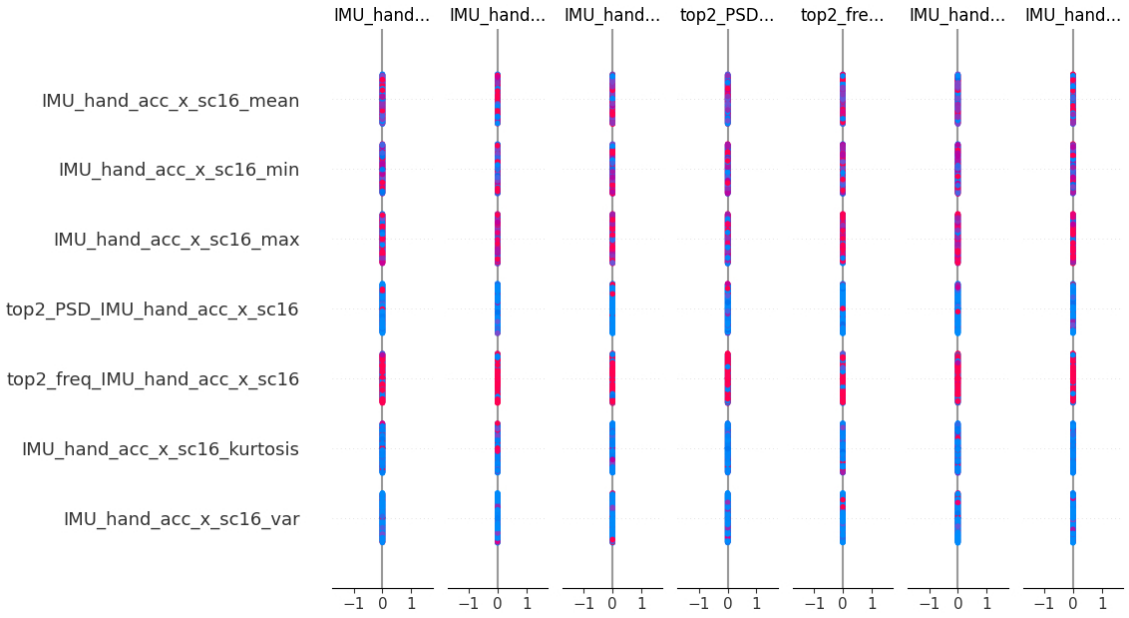
Table 4.10: Sensor Group Dropouts (Forced vs. Augmented)



The strongest drop from baseline can be noticed when the ankle sensor columns are zeroed out. However, in the augmented scenario, it is able to recover and even exceed the baseline accuracy. This denotes that even without the presence of all three groups of IMU sensors, the model is able to make a reliable prediction of user activity. It can also be noted that the heart rate values do not contribute anything meaningful to the activity prediction. Therefore, leveraging it as a proxy for exertion levels makes it functional.
Analyzing the LIME sensor grouped contributions, it can be noticed that the hand sensor group has less weight in comparison. Thus, it opens the doors for single IMU sensor-based testing on real-world implementations on end devices such as Smartphones or Smartwatches.

4.3 Interpretability of the Models

The top contributing features (original & extracted) for activity prediction from the SHAP interaction values are - the minimum IMU hand acceleration in x-axis, IMU hand acceleration kurtosis in x-axis, mean of IMU hand acceleration in x-axis, the top 2 frequencies of IMU hand acceleration in x-axis, IMU hand acceleration skew in x-axis & the top IMU hand acceleration frequency in x-axis.



Extracted features - top PSD filtered chest magnetometer in x-axis seems to have the highest positive LIME weight and top 3 frequencies of ankle orientation in y-axis seem to have the most negative LIME weights.

This denotes feature importance across the range; however, features on the x-axis generally had more positive contributions. So comparing dimensionality reduction may lead to faster processing on edge devices in the real world while retaining confident predictions.

Chapter 5

Conclusion

5.1 Discussion

The proposed methodology explores important challenges of HAR using the PAMAP2 dataset including aligning subject variability when performing particular activities, early recognition in fast-paced applications, reliability during sensor failures and the need for XAI in health-related applications and beyond. By combining traditional ML models, deep learning NN models, adjustments for different scenarios when multitask modeling with the added dimension of exertion levels and techniques for making the results explainable, practical insights for real-world usage were offered. The findings highlight that careful evaluation during pre-processing of IMU signals and reliable model selection are crucial for making HAR research effective in healthcare, fitness and assistive technologies of the present and the future. With the limited scopes of this study in mind, further work is necessary for tweaking the outcomes for end-device implementations at a larger scale.

5.2 Future Work

- Obtaining a HAR dataset with actual physical exertion labels from participant surveys.
- Using multi-modal fusion and exploring the weights of IMU signals vs. other modalities.
- Gathering Smartphone-based IMU signals only and implementing in an Activity Recognition application.
- Using Deep Neural Networks (DNNs) for Cross-Subject Adaptation.
- Exploring Sequence to Sequence Neural Networks for Activity Forecasting and comparing it with lightweight predictive models.
- Exploring privacy-preserving methods of IMU-based Activity Recognition.
- Exploring datasets with subjects from varying demographics.

Bibliography

[1] Y. Jiang, V. Hernandez, G. Venture, D. Kulić, and B. K. Chen, "A data-driven approach to predict fatigue in exercise based on motion data from wearable sensors or force plate," *Sensors*, vol. 21, no. 4, 2021, ISSN: 1424-8220. DOI: 10.3390/s21041499. [Online]. Available: https://www.mdpi.com/1424-8220/21/4/1499.

[2] I. E. Jaramillo, C. Chola, J.-G. Jeong, *et al.*, "Human activity prediction based on forecasted imu activity signals by sequence-to-sequence deep neural networks," *Sensors*, vol. 23, no. 14, 2023, ISSN: 1424-8220. DOI: 10.3390/s23146491. [Online]. Available: https://www.mdpi.com/1424-8220/23/14/6491.

[3] S. Seenath and M. Dharmaraj, "Conformer-based human activity recognition using inertial measurement units," *Sensors*, vol. 23, no. 17, 2023, ISSN: 1424-8220. DOI: 10.3390/s23177357. [Online]. Available: https://www.mdpi.com/1424-8220/23/17/7357.

[4] H. Nematallah and S. Rajan, "Adaptive hierarchical classification for human activity recognition using inertial measurement unit (imu) time-series data," *IEEE Access*, vol. 12, pp. 52 127–52 149, 2024. DOI: 10.1109/ACCESS.2024.3386351.

[5] Y. Xu, H. Cao, K. Mao, Z. Chen, L. Xie, and J. Yang, "Aligning correlation information for domain adaptation in action recognition," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 5, pp. 6767–6778, 2024. DOI: 10.1109/TNNLS.2022.3212909.

[6] A. M. Das, C. I. Tang, F. Kawsar, and M. Malekzadeh, "Primus: Pretraining imu encoders with multimodal self-supervision," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5. DOI: 10.1109/ICASSP49660.2025.10888874.

[7] S. Hwang, N. Kwon, D. Lee, *et al.*, "A multimodal fatigue detection system using semg and imu signals with a hybrid cnn-lstm-attention model," *Sensors*, vol. 25, no. 11, 2025, ISSN: 1424-8220. DOI: 10.3390/s25113309. [Online]. Available: https://www.mdpi.com/1424-8220/25/11/3309.

[8] X. Ye and K. I.-K. Wang, *Domain-adversarial anatomical graph networks for cross-user human activity recognition*, 2025. arXiv: 2505.06301 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2505.06301.