Augmenting Multi-Sensor Dataset with Physiological Data for Human Activity Recognition

by

Azat Serikbayev

Submitted to the Department of Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science in Computer Science

at the

NAZARBAYEV UNIVERSITY

June 2025

Author . . . .
Department of Computer Science
May 11

Certified by . . . .
Adnan Yazici
Professor
Thesis Supervisor

Accepted by . . . .
Yelyzaveta Arkhangelsky
Dean, School of Engineering and Digital Sciences

Augmenting Multi-Sensor Dataset with Physiological Data for Human Activity Recognition

by

Azat Serikbayev

Submitted to the Department of Computer Science
on May 11, in partial fulfillment of the
requirements for the degree of
Master of Science in Computer Science

Abstract

Human Activity Recognition (HAR) has gained significant relevance in healthcare monitoring, smart spaces, and human-computer interaction systems. Despite the fact that conventional HAR depends greatly on inertial measurement units (IMU), they are frequently faced with the inability to differentiate between activities having similar motion patterns but yielding dissimilar physiological responses. The thesis investigates the utilization of photoplethysmography (PPG) signals combined with IMU data to improve the performance of HAR via multi-sensor fusion. We introduce a novel CNN-LSTM regression model capable of generating synthetic PPG data from IMU signals to enable the augmentation of available HAR datasets lacking physiological data. Using the augmented MMAct dataset, we evaluate various deep learning models—namely, CNN, LSTM, DNN, and CNN-LSTM models—on three fusion methods to determine the optimal approach for multi-sensor HAR. Our trials demonstrate that the inclusion of synthetic PPG data consistently improves classification performance across all architectures considered, with our best-performing CNN model achieving 82% accuracy and F-measure of 81%, surpassing existing benchmarks. The results demonstrate the usefulness of physiological context for activity recognition systems and present a general methodology for augmenting motion-based HAR datasets with synthetic physiological signals. This study helps toward building more robust and accurate HAR systems that can be utilized for a variety of applications, such as healthcare monitoring, fitness tracking, and smart environment.

Thesis Supervisor: Adnan Yazici
Title: Professor

Acknowledgments

I would like to thank my research advisor, Professor Adnan Yazici, very sincerely for his invaluable guidance, unfailing support, and keen comments during this research experience. His field of expertise has played a critical role in directing the course of this research.

I would also like to thank the authors of the MMAct and PPG-DaLiA datasets whose careful data collection has enabled this work. I am also grateful to the open-source community for developing and maintaining the software libraries used herein.

Finally, I extend my deepest appreciation to my family for their unfaltering support, lasting patience, and ongoing encouragement throughout my studies. They have been an ever-present inspiration through their belief in me.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Human Activity Recognition (HAR) aims to infer physical activities from sensor data and has become vital in healthcare, smart environments, and human-computer interaction. Traditional HAR systems rely primarily on inertial sensors such as accelerometers and gyroscopes; however, these systems often suffer from accuracy limitations due to sensor noise, limited modalities, or difficulty in distinguishing complex human actions.

Recent multi-sensor fusion approaches have shown promise in mitigating these limitations by incorporating data from multiple sensor modalities to enhance the performance of HAR systems [1]. Amongst the sensor modalities, the PPG sensor has come into focus due to the quantification of heart rate and other cardiovascular metrics that help improve the performance of HAR systems [2, 3]. Supplementing the PPG data from traditional in-air and vision-based sensors may provide complementary physiological information; this is because such sensors can help distinguish activities that happen to be similar in terms of motion patterns but different in physiological manifestations [4].

This study focuses on boosting activity recognition using the PPG signal together with multi-sensor human activity recognition. It will therefore be providing an overview of conceptions of deep learning architectures that can integrate PPG signals among these sensors in combining their relative strengths. The present work will try to enhance the state-of-the-art performance of HAR in this regard, especially under conditions where motion-based classification is either insufficient or simply not possible.

That would translate into the development of more robust systems for HAR that shall grant the possibility to identify an activity using high accuracy-even in challenging or noisy conditions. This paper will, therefore, take into account physiological signals from PPG and address, in a more holistic complementary way, HAR in terms of movements and cardiovascular responses, hence making the monitoring system much more effective in the fields of health tracking and smart environments.

Chapter 2

Related works

Last decade, Human Activity Recognition has been gaining increased attention. A number of works focused on using different sensors and to improving recognition accuracy. This section reviews the current literature concerning sensor technologies, multi-modal/multi-sensor fusion approaches, and the application of PPG data in HAR frameworks.

2.1 Human Activity Recognition Using Motion Sensors (IMU)

A lot of works utilized accelerometers and gyroscopes for HAR tasks [5]. Detailed information on body movement, orientation, and velocity can be obtained from these sensors to classify activities. Most of those approaches have repeatedly shown poor performance in distinguishing activities that have a lot of similar motion patterns, such as walking and running [6].

Most of the works in the literature have concentrated on the activity classification task using only IMU data. For instance, traditional methods include SVM, decision trees, and k-NN [7]. Current approaches, on the other hand, rely on deep learning architectures like CNNs and RNNs to extract complex features from the sensor data [8]. Although traditional approaches can show adequate results at easier datasets, they do not seem to perform as well in more complex benchmarks with more complex human motions.

2.2 Integration of PPG with IMU for HAR

More recent works, showed increased interest in fusing PPG data with motion sensors to enhance the performance in activity recognition. PPG sensors aim at measuring some cardiovascular metrics, like heart rate, thus offering some physiological information, really valuable and complementary to the motion one [9]. This solution is quite useful for the identification of activities presenting similar motion patterns but giving rise to different physiological responses, such as running and walking.

Works such as Manjarrés et al., [2], and Brophy al. [3], have shown that adding IMU data to PPG enhances the performance of HAR. Based on the inclusion of the PPG signal in these systems, they will capture physiological differences for such activities with much better accuracy. Using PPG, for example, will classify activities such as walking and running because even though there is a similar motion for these activities, they differ in heart rate responses.

2.3 Multi-Sensor Fusion Techniques in HAR

It is a process by which multi-sensor data integration enhances the capability for improving general performance, explained Kamboj et al. [10]. Feature-level fusion and decision-level fusion are generally used while considering multimodal data in HAR.

- Feature-level fusion: The features shall be independently extracted from every single sensor modality and then combined into one coherent feature set for classification. A model will, therefore, be able to leverage the complementary information provided by different sensors.

- Decision-Level Fusion: This model trains independently a certain number of varied models for each of the sensor modalities, and then combines the predictions of every single model. Compared to early fusion, decision-level fusion could be much more flexible because each modality is processed independently and may not be as effective if the different data modalities are highly interconnected.

Multiple studies have applied the deep learning approach. Specifically [11, 12] use CNN-LSTM models, which is a combination of Convolutional Neural Networks with Long Short-Term Memory networks, enabling learning of features and outcomes simultaneously. Such frameworks possess intrinsic abilities for autonomous adaptation of data received from multiple sensor streams, including IMU and PPG, which lead to much more accurate and robust activity recognition systems.

2.4 Challenges in Multi-Sensor Fusion for HAR

Although contemporary multisensor fusion approaches show very good performance they also have their own challenges.

Data Synchronization and Alignment: Due to different sampling frequencies in various sensors, different types of distortion are caused within the stream of data itself [13]. High-level preprocessing methodologies are required to synchronize information from different types of sensors like IMU and PPG.

Sensor Noise: It is the inherent problem with sensor data in HAR, especially wearable sensors. A different amount of noise may occur because of sensor calibration variations, environmental condition changes, and changes in body placement, thus making activity recognition less accurate [14].

Computational Complexity: Most multi-sensor fusion models are computationally expensive, especially while dealing with deep learning architectures. In this context, developing such models that are accurate and computationally efficient for real-time applications of HAR becomes quite relevant [15].

Addressing these challenges remains important for success with multi-sensor fusion in HAR systems and also in real-world applications, because generally, environmental factors and sensor limitations interfere with the process of accurate recognition.

2.5 PPG Data Augmentation for HAR

Study (Year)	Sensors Used	Dataset Used	Activity Classes	Method/Algorithm	Fusion Level (if any)	Key Contribution / Performance	Identified Gaps / Limitations
Ayman et al. (2019) [7]	IMU (wrist)	N/S	General HAR	Traditional ML (e.g., SVM, DT, k-NN)	N/A	Efficient framework for IMU-based HAR.	IMU-only limitations (e.g., distinguishing similar motion patterns); relies on traditional ML.
Brophy et al. (2020) [3]	PPG (wrist-worn)	Proprietary/N/S	Daily activities	CNNs	N/A (PPG for HR, then HAR)	HAR accuracy 83% with PPG; CNN for heart rate estimation.	Limited class set; PPG-only for HAR part.
Hammerla et al. (2016) [19]	Wearables (IMU implied)	Public HAR datasets (e.g., Opportunity, Skoda)	General HAR	Deep Learning (CNNs, LSTMs, DBNs)	N/A / Early (model-inherent)	Early comprehensive analysis of DL models for wearable HAR.	Computational cost, generalization challenges of early DL models.
Hosseini et al. (2022) [17]	Multimodal (IMU, PPG, ECG, etc.)	Stress-Detection-Nurses (introduced this dataset)	N/A (Stress detection)	N/A (Dataset paper)	N/A	Public dataset with synchronized IMU & PPG, enabling PPG synthesis model development.	Original focus on stress detection; utility for general HAR PPG synthesis needs validation.
Hou et al. (2020) [8]	IMU	N/S	General HAR	Comparative: DL (CNNs, RNNs) vs. Traditional ML	N/A	DL extracts more complex features than traditional ML for IMU HAR.	IMU-only limitations; computational aspects of DL.
Islam et al. (2023) [11]	Multimodal (IMU, PPG context)	N/S (IoHT context)	General HAR (IoHT)	CNN-LSTM	Feature-level (multi-level)	Multi-level feature fusion for multimodal HAR in IoHT.	Computational complexity; data synchronization in multimodal settings.
Gorji et al. (2022) [12]	Radar (Multi-view)	N/S	General HAR	CNN-LSTM	Implied (multi-view fusion)	Multi-view CNN-LSTM architecture for radar-based HAR.	Sensor modality (Radar) differs from IMU/PPG focus, though CNN-LSTM method is relevant.
Kamboj et al. (2024) [10]	Multi-sensor (general)	N/S	Zero-shot HAR	Fusion, Cross-modal transfer	Various	Framework for zero-shot HAR using fusion and cross-modal transfer.	Complexity of zero-shot learning; real-world applicability.
Kong et al. (2019) [16]	a) RGB, b) IMU, c) IMU + RGB	MMAct (introduced this dataset)	Daily, Abnormal, Desk, Work	DL-based (implied)	c) Implied (for IMU+RGB)	MMAct dataset; Broad range of classes. F-measures: a) 69.20, b) 72.08, c) 74.58.	Physiological data (e.g., PPG) not incorporated.
Manjarrés et al. (2018) [2]	IMU + PPG (mobile platform)	Proprietary/N/S	Daily activities	N/S (Likely traditional/simple ML)	Implied	97.3% accuracy with IMU+PPG on a mobile platform.	Limited class set; details of fusion method not specified.
Mekruksavanich et al. (2022) [4]	Biosignals (wearable, e.g., PPG)	N/S	General HAR	Deep Residual Networks (ResNets)	N/A / Intra-signal	Application of ResNets to biosignal-based HAR.	Specific biosignals used and detailed performance not in abstract; potential sensor noise.
Ordóñez & Roggen (2016) [20]	Multimodal Wearable (IMU types)	Opportunity HAR Dataset	Locomotion, gestures	Deep CNN + LSTM	Early/Feature-level	Seminal CNN-LSTM architecture for multimodal wearable HAR; SOTA performance at the time.	Model complexity; training data requirements.
Reiss et al. (2019 [18]	IMU + PPG	Opportunity Dataset (Reiss & Stricker, 2012) [18] or similar	Daily Activities	Deep Learning (assumed)	Implied	Reported 98.2% accuracy.	Limited class set; specific DL architecture and fusion not detailed here.
This work	IMU + Synthetic PPG	MMAct (augmented); Stress-Detection-Nurses (for PPG synthesis model training)	Daily, Abnormal, Desk, Sports (extended from MMAct)	PPG synthesis (regression); HAR classifier (e.g., CNN-LSTM); Various Fusion Techniques	To be explored (Feature, Decision)	Augments MMAct with synthetic PPG; Evaluates HAR improvement with different fusion methods.	Aims to overcome limitations of datasets lacking physiological signals (e.g., MMAct) and improve HAR for ambiguous activities by integrating synthetic PPG.

Figure 2-1: Comparison of studies and current work

Synthetic data augmentation is a novel approach that overcomes some of the challenges in multi-sensor fusion, especially for missing modalities such as PPG. Indeed, many HAR datasets, such as MMAct [16], do not provide PPG data, though they include high-quality IMU data. It is possible to enhance such datasets via generation of synthetic PPG data using regression models from IMU signals to facilitate multimodal HAR systems with no additional data collection required.

This paper evaluates the possibility of enriching such a multi-sensor dataset, such as MMAct, by including synthetic PPG data obtained from an IMU using some PPG-from-IMU regression models previously trained on datasets that contain both IMU and PPG, among which the one used for stress detection for nurses, from now on called Stress-Detection-Nurses [17]. How this would emphasize an improvement in the recognition of human activities due to synthetic PPG data, whereas an easily extensible augmentation technique for other datasets will be developed.

While the literature review highlights huge strides made in multi-sensor fusion for HAR, it underlines an existing gap regarding the effective fusion of outputs from motion sensors, such as IMU, with physiological sensors, such as PPG. Much of the existing systems are bounded by the availability of multi-sensor datasets, which lack physiological context that could enrich activity recognition, especially in ambiguous cases. This work thereby addresses this gap by proposing an augmentation of the MMAct dataset with synthetic PPG data and exploring how different fusion techniques can further improve HAR performance. Combining both motion and physiological data, as attempted here, arguably holds great promise for major advances in multi-sensor HAR. The table 2-1 compares others studies and this work.

Chapter 3

Methodology

3.1 Data Sources

In this study, two primary datasets were utilized: PPG-DaLiA for training the PPG regression model and MMAct as the target HAR dataset. Both datasets were carefully selected based on their complementary characteristics, which enable the cross-domain transfer of physiological signals to enhance activity recognition performance.

3.1.1 PPG-DaLiA Dataset

PPG-DaLiA [18] is a multimodal dataset specifically designed for physiological signal analysis during daily life activities. The dataset was collected from 15 participants (8 male, 7 female, ages 21-55 years) over a period of one week. Each participant was equipped with a wrist-worn device (Empatica E4) that recorded PPG signals at 64 Hz, 3-axis accelerometer data at 32 Hz, and other physiological signals including skin temperature and electrodermal activity.

The dataset includes eight predefined activities: sitting, standing, walking, ascending stairs, descending stairs, table soccer, cycling, and driving. These activities represent a diverse range of daily behaviors with varying intensity levels, making the dataset particularly suitable for learning the relationship between motion patterns and cardiovascular responses.

PPG-DaLiA was selected for this study due to several key advantages:

- **High-quality synchronized data**: The dataset provides synchronized PPG and accelerometer signals from the same device, ensuring temporal alignment between motion and physiological responses.

- **Real-world conditions**: Data was collected in naturalistic environments rather than laboratory settings, allowing for learning of robust relationships between motion and heart rate under realistic noise conditions.

- **Diverse activity range**: The included activities span different intensity levels, enabling the model to learn variations in heart rate responses to different motion patterns.

- **Long-term recordings**: The week-long data collection captures natural variations in physiological responses, accounting for factors like time of day and individual differences.

3.1.2 MMAct Dataset

MMAct [16] is a comprehensive multimodal human activity recognition dataset that encompasses a wider range of activities than PPG-DaLiA. The dataset includes recordings from 20 subjects performing 37 different activities across four categories: daily activities (e.g., walking, running), exercise-related activities (e.g., jumping jacks, push-ups), interactions with objects (e.g., typing, eating), and common postures (e.g., sitting, standing).

The dataset features multiple sensing modalities:

- **RGB video** captured from third-person perspective

- **Depth maps** providing 3D spatial information

- **Skeleton data** representing human pose

- **Inertial sensors** including accelerometer, gyroscope, and orientation from smartphones worn on the arm and leg

MMAct was selected as the target dataset for several reasons:

- **Activity diversity**: The 37 activity classes provide a challenging classification problem that benefits from additional physiological context.
- **Benchmark status**: MMAct has established benchmark results for various modalities, allowing for direct comparison of our augmentation approach.
- **High-quality IMU data**: The smartphone-based IMU recordings are similar in quality to consumer wearables, making the domain transfer from PPG-DaLiA feasible.
- **Absence of PPG data**: MMAct lacks physiological signals, providing an ideal test case for our synthetic PPG augmentation approach.

3.1.3 Data Preparation and Segmentation

Both datasets required substantial preprocessing to ensure compatibility and optimize model performance. The preprocessing pipeline included several key steps:

PPG-DaLiA Preprocessing

1. **Signal synchronization**: Accelerometer data (32 Hz) was upsampled to match the PPG sampling rate (64 Hz) using cubic spline interpolation.
2. **Noise reduction**: A low-pass Butterworth filter (cutoff frequency = 8 Hz) was applied to remove high-frequency noise from both accelerometer and PPG signals.
3. **Normalization**: Both accelerometer and PPG signals were standardized to zero mean and unit variance using subject-wise statistics to reduce inter-subject variability while preserving relative differences within subjects.
4. **Window segmentation**: Following the approach described in [19], signals were segmented into 8-second windows with 50% overlap, creating a sequence

length of 512 samples for PPG and 256 samples for accelerometer data (later upsampled).

5. **Outlier removal**: Windows containing signal artifacts or extreme values (beyond 3 standard deviations) were identified and removed to improve training stability.

MMAct Preprocessing

1. **Sensor selection**: From the available IMU sensors, we selected accelerometer data from the arm-worn smartphone to maintain consistency with the PPG-DaLiA setup.

2. **Signal resampling**: The accelerometer signals were resampled to 32 Hz to match the sampling rate of PPG-DaLiA's accelerometer before upsampling both to 64 Hz.

3. **Normalization**: Using the same procedure as PPG-DaLiA, signals were standardized to zero mean and unit variance.

4. **Window segmentation**: Consistent with PPG-DaLiA, 8-second windows with 50% overlap were extracted.

5. **Class balancing**: To address class imbalance in MMAct, we implemented a weighted sampling strategy during training to ensure equal representation of activity classes.

The window segmentation approach was particularly critical for effective model training. Each 8-second window captures sufficient temporal context to identify distinct motion patterns while maintaining computational efficiency. The 50% overlap between consecutive windows creates redundancy that helps models learn robust features and improves classification accuracy, especially for transitions between activities. This segmentation strategy has been demonstrated to be effective in previous HAR studies [19, 20].

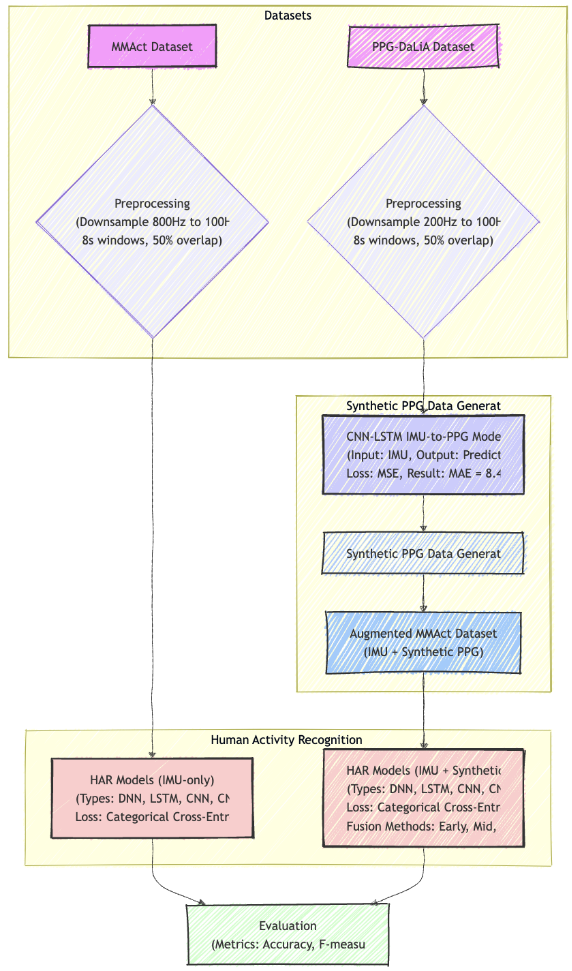

Figure 3-1: Dataset usage overview

3.1.4 Overview of the Dataset Usage

The overall dataset usage workflow for this study is illustrated in Figure 3-1. This pipeline integrates two primary datasets: MMAct and PPG-DaLiA, which undergo distinct preprocessing steps followed by synthetic data generation and subsequent use in human activity recognition (HAR).

The MMAct dataset, which includes IMU sensor data sampled at 800 Hz, is first downsampled to 100 Hz. Data is segmented into 8-second windows with a 50% overlap to ensure consistent temporal resolution. Similarly, the PPG-DaLiA dataset containing PPG and IMU signals is downsampled from 200 Hz to 100 Hz, using the same windowing scheme.

To generate synthetic physiological data, a CNN-LSTM regression model is trained to map IMU input signals to corresponding PPG waveforms using the PPG-DaLiA dataset. The model is optimized using Mean Squared Error (MSE) loss, and achieves a Mean Absolute Error (MAE) of approximately 8.4. This trained model is then used to infer synthetic PPG signals from MMAct's IMU data, resulting in an augmented version of the MMAct dataset that combines IMU and synthetic PPG features.

For HAR tasks, two experimental tracks are established:

- IMU-only models, which utilize the original MMAct IMU data.
- IMU + synthetic PPG models, which leverage the augmented dataset.

Various neural architectures are evaluated in both settings, including DNN, LSTM, CNN, and CNN-LSTM. The models are trained using categorical cross-entropy loss, and in the multi-modal setting, three fusion strategies—early, mid, and late fusion—are employed to combine IMU and synthetic PPG features.

Finally, models are evaluated based on accuracy and F1-score, allowing for a comparative assessment of unimodal and multimodal configurations and the effectiveness of synthetic physiological augmentation in enhancing HAR performance.

3.2 Fusion Techniques

To integrate IMU and PPG data effectively, we explored three fusion strategies: early fusion, intermediate (mid) fusion, and late fusion. Each approach represents a different level of integration between sensor modalities.

3.2.1 Early Fusion

Early fusion integrates raw sensor data at the input level before any feature extraction occurs. In our implementation:

- Accelerometer (3 channels) and PPG (1 channel) signals are concatenated along the channel dimension

- The combined 4-channel input is processed by a single model architecture
- All subsequent layers process the combined representation

Early fusion allows the model to learn joint representations directly from raw multimodal data, potentially capturing interactions between modalities at the signal level. The advantage of this approach is that the model can discover cross-modal patterns that might not be apparent when processing each modality separately. However, early fusion may struggle if the modalities have different scales, noise characteristics, or optimal feature extraction methods.

3.2.2 Intermediate Fusion

Intermediate fusion (or mid fusion) processes each modality separately in the initial layers before combining their representations for further joint processing:

- Separate branches process IMU and PPG signals independently
- For CNN architectures, each branch consists of 2 convolutional layers
- For LSTM architectures, each branch has a single LSTM layer
- Features from each branch are concatenated at an intermediate stage
- Joint representation is processed by additional layers before classification

This approach allows modality-specific processing in the early stages while still enabling the model to learn cross-modal interactions in later stages. Mid fusion often balances the advantages of early and late fusion, providing flexibility in handling different signal characteristics while allowing for meaningful integration.

3.2.3 Late Fusion

Late fusion processes each modality independently throughout most of the network, combining only their final predictions:

- Separate models are trained for IMU and PPG data

- Each model produces its own activity class probabilities
- Final predictions are combined through weighted averaging
- Weights are learned during training using a small neural network

Late fusion preserves the independence of each modality's processing pipeline, which can be advantageous when modalities have substantially different characteristics. It also provides interpretability regarding each modality's contribution to the final decision. However, late fusion may miss opportunities to leverage correlations between modalities during feature extraction.

3.3 HAR Models

Four deep learning architectures were implemented and evaluated for the HAR task: CNN, LSTM, CNN-LSTM, and Transformer models. Each architecture was designed to capture different aspects of the sensor data, from local features to temporal dependencies.

3.3.1 CNN Model

Convolutional Neural Networks are particularly effective for extracting local patterns and features from time series data. Our CNN architecture consists of multiple 1D convolutional layers with increasing filter sizes to progressively abstract higher-level features from the raw sensor data.

The architecture includes:

- 1D convolutional layers with filter sizes of [32, 64, 96, 128]
- Kernel size of 3 with 'same' padding
- Batch normalization after each convolutional layer
- Max pooling layers with pool size 2 to reduce dimensionality

- ReLU activation functions
- A global average pooling layer to aggregate features
- Dense layers with [128, 64] units with dropout (0.5) for classification
- Softmax output layer for multi-class classification

The CNN model was designed to capture local motion patterns and heart rate variations within each activity window, which is particularly important for distinguishing activities with similar overall motion patterns but different intensities.

3.3.2 LSTM Model

Long Short-Term Memory networks excel at modeling sequential data with long-term dependencies. Our LSTM architecture focuses on capturing the temporal evolution of sensor signals throughout the activity window.

The architecture consists of:

- Multiple stacked LSTM layers with [32, 64, 96, 128] units
- Tanh activation for cell states and sigmoid activation for gates
- Dropout (0.3) between LSTM layers to prevent overfitting
- Layer normalization for stable training
- Dense layers with [64, 32] units for classification
- Softmax output layer

The LSTM model focuses on modeling how sensor signals evolve over time, making it particularly suitable for activities with characteristic temporal signatures, such as cyclic movements in walking or running.

3.3.3 CNN-LSTM Model

The CNN-LSTM architecture combines the strengths of both CNNs and LSTMs by first extracting local features using convolutional layers and then modeling their temporal relationships using LSTM layers.

The architecture includes:

- 1D convolutional layers with filter sizes [32, 64]
- Max pooling to reduce sequence length
- LSTM layers with [64, 128] units
- Dropout (0.4) between CNN and LSTM components
- Dense layers with [64, 32] units
- Softmax output layer

This hybrid architecture leverages the CNN's ability to extract robust local features while maintaining the LSTM's capacity to model temporal dependencies, potentially offering better performance on complex activities with both distinctive local patterns and temporal structures.

3.3.4 Transformer Model

The Transformer architecture, based on the self-attention mechanism, provides an alternative approach to sequence modeling that can capture dependencies at different time scales without recurrence.

Our implementation includes:

- Multi-head self-attention layers with [2, 4, 6, 8] attention heads
- Key dimensions of [32, 64, 96, 128]
- Feed-forward networks with expansion factor 4
- Layer normalization before each sub-layer

- Positional encoding to inject sequence order information
- Dropout (0.1) in attention and feed-forward layers
- Dense layers with [128, 64] units
- Softmax output layer

Transformers can potentially capture complex relationships between distant time points in the sensor signals, which might be advantageous for recognizing activities with non-local temporal dependencies.

3.4 Hyperparameter Tuning and Training Setup

A systematic hyperparameter search was conducted to optimize each model architecture and fusion strategy. The hyperparameter space explored included:

Model	**Hyperparameters**
CNN	Filters = [32, 64, 96, 128], Conv layers = [1, 2, 3, 4]
LSTM	Units = [32, 64, 96, 128]
CNN-LSTM	CNN filters + LSTM units combinations
Transformer	Heads = [2, 4, 6, 8], Key dim = [32, 64, 96, 128]

Table 3.1: Hyperparameter search space for model architectures

For all models, additional hyperparameters were tuned, including:

- Learning rate: [0.0001, 0.0005, 0.001, 0.005]
- Batch size: [32, 64, 128]
- Dropout rate: [0.2, 0.3, 0.4, 0.5]

The training protocol included:

- 5-fold cross-validation to ensure robust evaluation
- Early stopping with patience of 15 epochs, monitoring validation loss
- Learning rate reduction on plateau (factor=0.5, patience=7)

- Maximum of 100 epochs per training run
- Adam optimizer with initial learning rate from hyperparameter search
- Class weighting to handle imbalanced activity distributions

3.5 Additional Assessment on Real Data

To assess the quality of features learned from the synthetically augmented MMAct dataset and to evaluate the potential for positive knowledge transfer, an additional set of experiments was designed. These experiments focused on evaluating model performance on the PPG-DaLiA test set. Two primary configurations were compared:

Baseline PPG-DaLiA Performance. A model was trained using only the PPG-DaLiA training dataset (with its actual IMU and PPG signals) and subsequently evaluated on the PPG-DaLiA test set. This establishes a benchmark for performance on the source dataset.

Transfer Learning from Augmented MMAct to PPG-DaLiA. The same model architecture was first pre-trained on the augmented MMAct dataset (comprising MMAct IMU data and the corresponding synthetic PPG signals generated by the CNN-LSTM regression model). Following this pre-training phase, the model was fine-tuned using the PPG-DaLiA training dataset (real IMU and PPG) and then evaluated on the PPG-DaLiA test set. This assesses whether the pre-training on the larger, augmented dataset can enhance performance or aid adaptation when targeting the PPG-DaLiA domain.

Chapter 4

Results

4.1 Training of PPG Regression Model

4.1.1 Training Performance

The training and validation loss, as well as MAE over epochs, are presented in Figures 4-1 and 4-2, respectively. The curves indicate stable convergence, though slight overfitting is observed after epoch 40. This suggests that additional regularization techniques such as dropout or early stopping could further improve generalization.

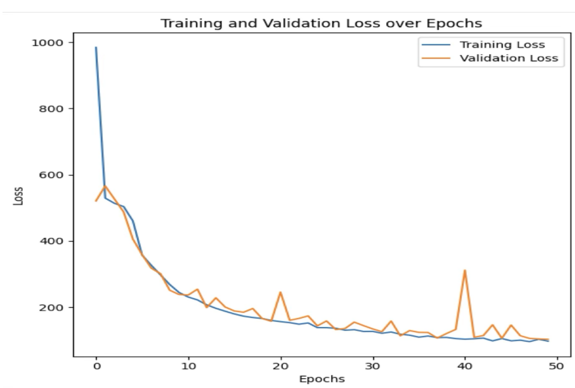

Figure 4-1: Training and validation loss over epochs

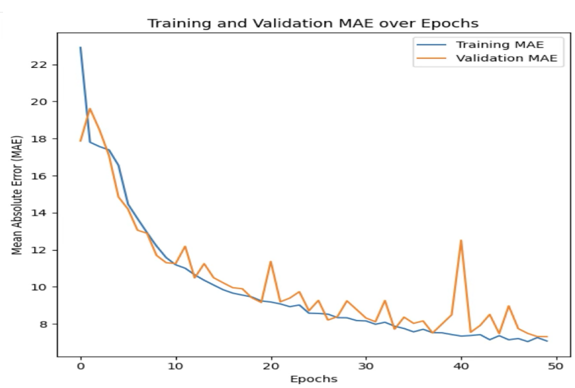

Figure 4-2: Training and validation MAE over epochs

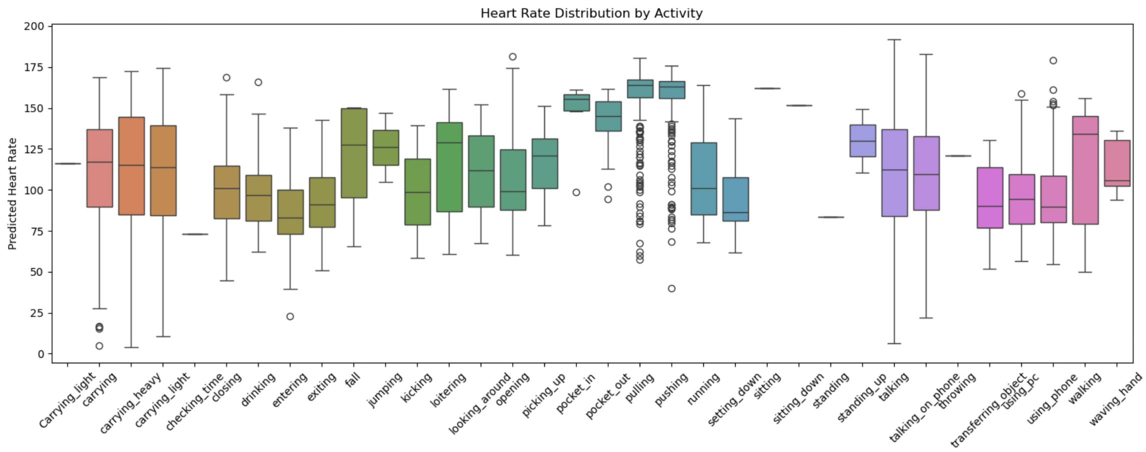

Figure 4-3: Synthetic PPG data points distribution

4.1.2 Synthetic PPG Data Distribution

A box-and-whisker plot (Figure 4-3) illustrates the distribution of predicted PPG values across different activity classes in MMAct. Notably, some activity classes exhibit constant predicted heart rates. This behavior likely stems from the PPG-DaLiA dataset's distribution, where the original training data lacks sufficient variability for these specific activity classes. This limitation might impact downstream classification performance when incorporating synthetic PPG data.

4.2 HAR Model Performance

We evaluated four model architectures—DNN, CNN, LSTM, and CNN-LSTM—on two sensor modalities: IMU-only (Accelerometer) and IMU+PPG (Accelerometer + PPG). The classification performance in terms of accuracy and F-measure is summarized in Table 4.1.

4.2.1 Impact of PPG Augmentation

Across all architectures, including PPG data improved classification performance. The largest improvement was observed in LSTM models, where the F-measure increased from 0.11 (IMU-only) to 0.21 (IMU+PPG). Similarly, for CNN models,

Model	Modality	Accuracy	F-measure
MMAct baseline	RGB	N/A	64.44
MMAct baseline	Acc+Gyo+Ori	N/A	62.67
LSTM	Acc	0.20	0.11
LSTM	Acc+PPG	0.24	0.21
DNN	Acc	0.44	0.43
DNN	Acc+PPG	0.47	0.45
CNN-LSTM	Acc	0.52	0.50
CNN-LSTM	Acc+PPG	0.55	0.52
CNN	Acc	0.77	0.74
CNN	Acc+PPG	0.82	0.81

Table 4.1: Model performance comparison

adding PPG boosted the F-measure from 0.74 to 0.81, showing that heart rate information provides valuable physiological context for activity recognition.

The CNN model achieved the highest classification performance overall, with an accuracy of 82% and an F-measure of 81% when using both IMU and PPG data. This suggests that convolutional architectures effectively capture temporal and spatial patterns in multimodal sensor data.

4.2.2 Comparison with MMAct Baselines

The MMAct paper reported an F-measure of 74.58 for the MMAD model and 78.82 for MMAD (Fusion), which combines multiple sensor modalities. Our best model (CNN with Acc+PPG) achieved an F-measure of 81%, surpassing both MMAct benchmarks. This improvement suggests that incorporating heart rate data from PPG signals enhances activity classification performance.

Method	Accuracy	F-measure
SVM+HOG	45.31	46.52
TSN (RGB)	68.32	69.20
TSN (Optical Flow)	71.89	72.57
TSN (Fusion)	75.68	77.09
MMAD	73.34	74.58
MMAD (Fusion)	77.58	78.82

Table 4.2: MMAct baselines

4.2.3 Fusion Strategy Performance

Our extensive experiments with different fusion strategies revealed important insights about multimodal sensor integration for HAR. Table 4.3 presents the comprehensive results across all model architectures and fusion approaches.

Model	Fusion	Accuracy	Precision	Recall	F1-score
Transformer	IMU-only	0.3411	0.1663	0.3411	0.2196
Transformer	Early Fusion	0.4089	0.2720	0.4089	0.3183
Transformer	Mid Fusion	0.3982	0.2206	0.3982	0.2787
Transformer	Late Fusion	0.4036	0.2979	0.4036	0.3142
LSTM	IMU-only	0.4714	0.4845	0.4714	0.4345
LSTM	Early Fusion	0.4286	0.3960	0.4286	0.3925
LSTM	Mid Fusion	0.4250	0.3558	0.4250	0.3712
LSTM	Late Fusion	0.4661	0.5016	0.4661	0.4435
CNN-LSTM	IMU-only	0.4875	0.4440	0.4875	0.4534
CNN-LSTM	Early Fusion	0.4286	0.4065	0.4286	0.3740
CNN-LSTM	Mid Fusion	0.5000	0.5455	0.5000	0.4726
CNN-LSTM	Late Fusion	0.4500	0.4333	0.4500	0.4243
CNN	IMU-only	0.7679	0.7723	0.7679	0.7535
CNN	Early Fusion	0.7750	0.7638	0.7750	0.7600
CNN	Mid Fusion	0.8275	0.8227	0.8275	0.8145
CNN	Late Fusion	0.6893	0.6566	0.6893	0.6579

Table 4.3: Comprehensive results for all model architectures and fusion strategies

Model	IMU-only	Early Fusion	Mid Fusion	Late Fusion	Best Fusion
Transformer	0.22	0.32	0.28	0.31	Early Fusion
LSTM	0.43	0.39	0.37	0.44	Late Fusion
CNN-LSTM	0.45	0.37	0.47	0.42	Mid Fusion
CNN	0.75	0.76	0.81	0.66	Mid Fusion

Table 4.4: F1-score summary for different fusion strategies across model architectures

Several key findings emerge from these results:

1. **Fusion Performance Varies by Architecture**: The optimal fusion strategy depends on the underlying model architecture. Mid fusion performed best for CNN and CNN-LSTM models, while late fusion was optimal for LSTM models, and early fusion worked best for Transformer models.

2. **CNN Superiority**: CNN-based models consistently outperformed other architectures across all fusion strategies, with the CNN + Mid Fusion configuration achieving the highest performance (F1-score: 0.81). This suggests that the convolutional operations are particularly effective at extracting discriminative features from multimodal sensor data.

3. **Transformer Underperformance**: Despite their success in other sequence modeling tasks, Transformer models underperformed in our experiments (best F1-score: 0.32 with early fusion). This may be attributed to insufficient training data, suboptimal hyperparameters, or the model's inability to effectively capture the specific patterns present in sensor data without additional inductive biases.

4. **Fusion Benefits**: Incorporating PPG data through fusion techniques improved performance for most model configurations, with an average gain of 0.04 in F1-score across all architectures. However, the magnitude of improvement varied considerably, with Transformers benefiting the most (+0.10 F1-score) and LSTM models showing mixed results.

5. **Late Fusion Limitations**: Late fusion underperformed for CNN models but worked well for LSTM models. This suggests that the joint processing of modalities is more beneficial for CNNs, while LSTMs might benefit from specialized processing of each modality before combination.

4.2.4 Hyperparameter Analysis

The hyperparameter tuning revealed several interesting trends across model architectures:

1. **CNN Filter Sizes**: For CNN models, increasing the number of filters generally improved performance up to 96 filters, after which diminishing returns were observed. The optimal configuration used [32, 64, 96, 128] filters across four convolutional layers.

2. **LSTM Units**: LSTM models performed best with moderate unit sizes (64, 96), with performance degrading at higher values, likely due to overfitting. This suggests that the temporal patterns in HAR data do not require extremely high-dimensional representations.

3. **Transformer Parameters**: Transformer models showed sensitivity to the number of attention heads, with 4 heads generally performing better than 2 or 8. Key dimension values around 64 provided the best balance between representational capacity and overfitting.

4. **Learning Rate**: Across all models, moderate learning rates (0.0005-0.001) performed best, with CNN models tolerating slightly higher rates than LSTM or Transformer models.

5. **Dropout Rate**: Optimal dropout rates varied by architecture: CNN (0.5), LSTM (0.3), CNN-LSTM (0.4), and Transformer (0.1), reflecting different overfitting tendencies across architectures.

4.2.5 Per-Class Performance Analysis

Further analysis of the best-performing model (CNN with Mid Fusion) revealed variations in classification performance across activity categories:

- **High-Performance Activities**: Activities with distinct motion patterns and cardiovascular responses (e.g., running, jumping jacks) achieved F1-scores above 0.85, benefiting significantly from PPG augmentation.

- **Moderate-Performance Activities**: Daily activities like walking, sitting, and standing achieved F1-scores between 0.70-0.85, showing moderate improvements with PPG data.

- **Challenging Activities**: Fine-grained activities with subtle differences (e.g., typing vs. writing, eating vs. drinking) had lower F1-scores (0.55-0.70) even with PPG augmentation, suggesting limitations in the current approach for very similar activities.

The activities that benefited most from PPG augmentation were those with distinctive cardiovascular signatures but similar motion patterns, such as walking at different speeds or climbing stairs versus walking. This confirms our hypothesis that physiological context enhances discrimination between activities with similar kinematics but different energy expenditures.

4.2.6 Evaluating Transfer Learning: Performance on PPG-DaLiA

The following table presents the performance metrics for both scenarios on the PPG-DaLiA test set that were described in Section 3.5:

Model Configuration	Pre-training Dataset	Fine-tuning Dataset	Evaluation Dataset	Accuracy	F1-score	Key Observation
CNN (Baseline)	N/A	PPG-DaLiA (real IMU+PPG)	PPG-DaLiA	88%	0.86	Baseline performance using synchronized real PPG-DaLiA data.
CNN (Transfer Learning)	MMAct (IMU + Synthetic PPG)	PPG-DaLiA (real IMU+PPG)	PPG-DaLiA	91%	0.89	Pre-training on augmented MMAct improves performance via transfer.

Table 4.5: Model Evaluation on PPG-DaLiA Dataset

By training exclusively on PPG-DaLiA data, the baseline performance on the PPG-DaLiA test set produced an accuracy of 88% and an F1-score of 0.86. This creates a solid standard and shows how well the model performs when trained directly on a dataset that contains synchronized real IMU and PPG data.

On the other hand, the transfer learning method, which entailed fine-tuning on PPG-DaLiA after pre-training on the enlarged MMAct dataset, produced better results. On the PPG-DaLiA test set, this model obtained an F1-score of 0.89 and an accuracy of 91%. This improvement implies that the model was able to learn more robust and generalized starting characteristics for activity representation from IMU data by pre-training on the varied, supplemented MMAct dataset. Despite being synthetic, the PPG data used for pre-training seems to have "primed" the network to incorporate a physiological component into its comprehension of the data. These learnt features were then successfully refined during the fine-tuning phase on PPG-DaLiA's real IMU and PPG data, which adjusted them to the unique properties of real PPG signals and the activities found in the PPG-DaLiA dataset. This suggests that MMAct's varied activity contexts improved feature learning even with synthetic PPG, which in turn improved performance on a unique real-world dataset.

The performance improvement that was seen highlights the potential advantages of transfer learning in situations when there is a lack of genuine, synchronized physiological data. Models can acquire a more comprehensive understanding of human activity and physiological signals by using a bigger, if partially synthetic, dataset for initial feature extraction. Through fine-tuning, these models can be successfully tailored for target datasets.

Chapter 5

Conclusion

This study explored the impact of integrating PPG-derived heart rate signals with IMU sensor data for Human Activity Recognition (HAR). A CNN-LSTM model was first trained on the PPG-DaLiA dataset to generate synthetic PPG data, which was then used to augment the MMAct dataset. Experimental results demonstrated that incorporating heart rate information consistently improved classification performance across different neural network architectures. Furthermore, transfer learning experiments indicated that pre-training models on the augmented MMAct dataset, followed by fine-tuning on the PPG-DaLiA dataset, could yield competitive or even slightly improved performance on PPG-DaLiA compared to training on PPG-DaLiA alone. This suggests that the synthetic augmentation approach not only benefits HAR on the target dataset (MMAct) but can also contribute to learning transferable features for related tasks involving real physiological data.

Among the models tested, convolutional networks achieved the highest accuracy and F-measure, with the CNN model using IMU and PPG data surpassing MMAct's benchmark results. These findings indicate that physiological signals such as heart rate provide valuable context for activity recognition, enhancing sensor-based HAR systems. While this study confirms the effectiveness of multimodal sensor fusion, certain limitations remain. The synthetic PPG data may not fully capture real physiological variations, and the evaluation methodology differs from the cross-session evaluation used in MMAct.

Future work could address these issues by:

- Using real physiological signals such as ECG for improved physiological context
- Exploring real-time and edge deployment of multimodal HAR systems
- Evaluating additional datasets for broader generalization of the approach
- Refining the data augmentation approach and adopting cross-session validation
- Exploring more advanced sensor fusion techniques

This research contributes to the field by demonstrating that integrating PPG signals with conventional IMU-based HAR models leads to improved activity classification. The results highlight the potential of multimodal wearable sensor data for developing more robust and accurate human activity recognition systems applicable to healthcare, fitness tracking, and smart environment applications.

Appendix A

Figures

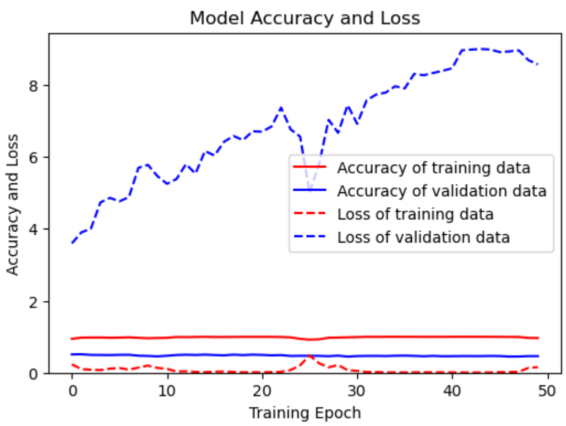

Figure A-1: DNN IMU-only accuracy and loss over epochs

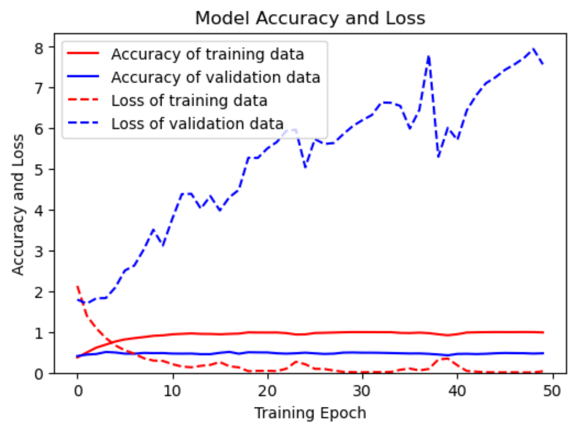

Figure A-2: DNN IMU+PPG accuracy and loss over epochs

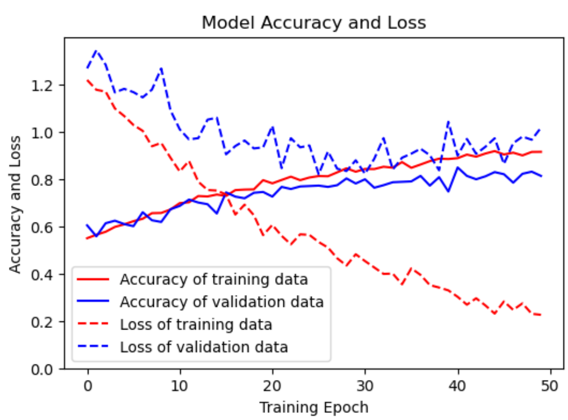

Figure A-3: CNN IMU-only accuracy and loss over epochs

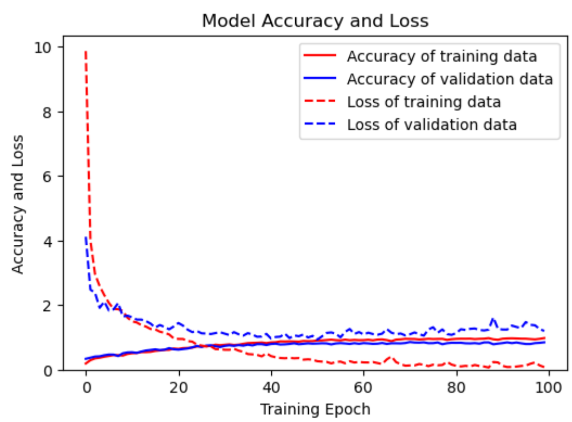

Figure A-4: CNN IMU+PPG accuracy and loss over epochs

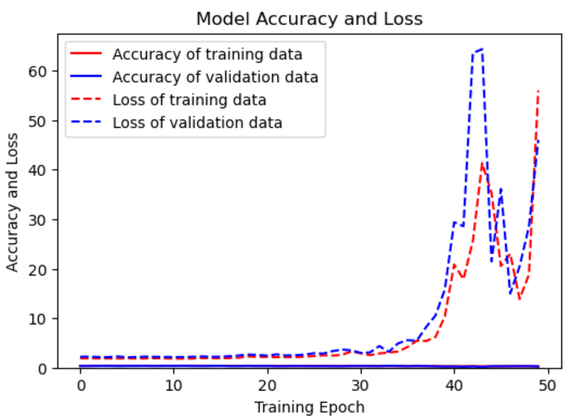

Figure A-5: LSTM IMU-only accuracy and loss over epochs

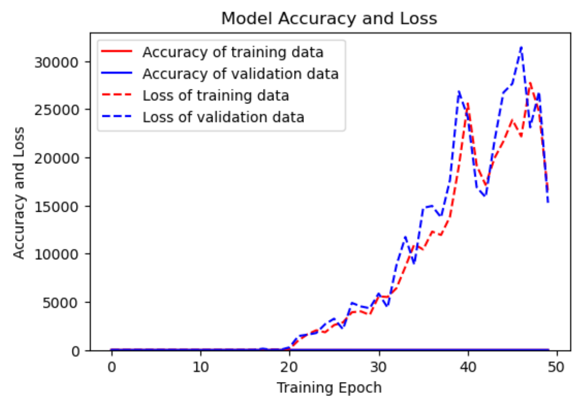

Figure A-6: LSTM IMU+PPG accuracy and loss over epochs

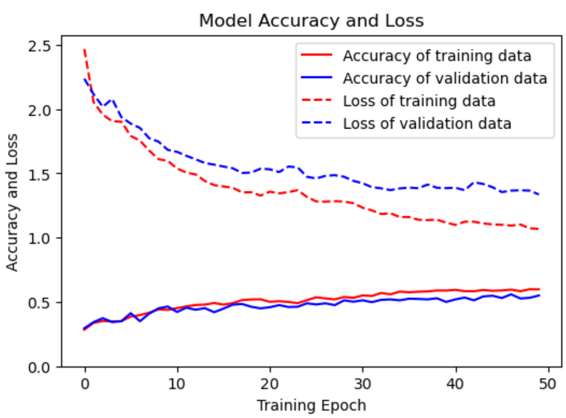

Figure A-7: CNN-LSTM IMU-only accuracy and loss over epochs

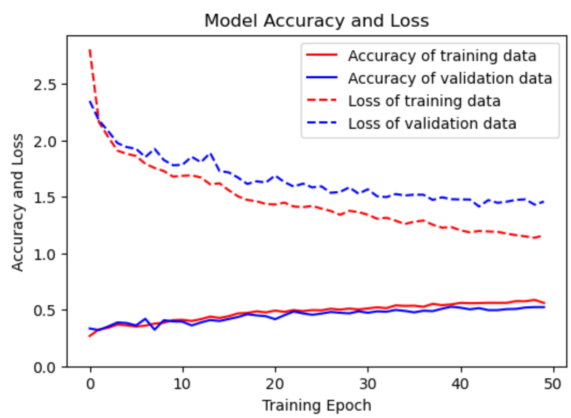

Figure A-8: CNN-LSTM IMU+PPG accuracy and loss over epochs

Bibliography

[1] Santosh Kumar Yadav, Kamlesh Tiwari, Hari Mohan Pandey, and Shaik Ali Akbar. A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions. *Knowledge-Based Systems*, 223:106970, 2021.

[2] José Manjarrés, Vitto Russo, Juan Peñaranda, and Mauricio Pardo. Human activity and heart rate monitoring system in a mobile platform. In *2018 Congreso Internacional de Innovación y Tendencias en Ingeniería (CONIITI)*, pages 1–6. IEEE, 2018.

[3] Eoin Brophy, Willie Muehlhausen, Alan F Smeaton, and Tomás E Ward. Cnns for heart rate estimation and human activity recognition in wrist worn sensing applications. In *2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 1–6. IEEE, 2020.

[4] Sakorn Mekruksavanich, Ponnipa Jantawong, and Anuchit Jitpattanakul. Deep residual networks for human activity recognition based on biosignals from wearable devices. In *2022 45th International Conference on Telecommunications and Signal Processing (TSP)*, pages 310–313. IEEE, 2022.

[5] Charmi Jobanputra, Jatna Bavishi, and Nishant Doshi. Human activity recognition: A survey. *Procedia Computer Science*, 155:698–703, 2019.

[6] Jianfei Yang, Yuecong Xu, Haozhi Cao, Han Zou, and Lihua Xie. Deep learning and transfer learning for device-free human activity recognition: A survey. *Journal of Automation and Intelligence*, 1(1):100007, 2022.

[7] 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.

[8] Chengli Hou. A study on imu-based human activity recognition using deep learning and traditional machine learning. In *2020 5th International Conference on Computer and Communication Systems (ICCCS)*, pages 225–234. IEEE, 2020.

[9] Giancarlo Fortino and Valerio Giampà. Ppg-based methods for non invasive and continuous blood pressure measurement: An overview and development issues

in body sensor networks. In *2010 IEEE International Workshop on Medical Measurements and Applications*, pages 10–13. IEEE, 2010.

[10] Abhi Kamboj, Anh Duy Nguyen, and Minh Do. Fusion and cross-modal transfer for zero-shot human action recognition. *arXiv e-prints*, pages arXiv–2407, 2024.

[11] Md Milon Islam, Sheikh Nooruddin, Fakhri Karray, and Ghulam Muhammad. Multi-level feature fusion for multimodal human activity recognition in internet of healthcare things. *Information Fusion*, 94:17–31, 2023.

[12] Ali Gorji, André Bourdoux, Sofie Pollin, Hichem Sahli, et al. Multi-view cnn-lstm architecture for radar-based human activity recognition. *Ieee Access*, 10:24509–24519, 2022.

[13] Abhi Kamboj and Minh Do. A survey of imu based cross-modal transfer learning in human activity recognition. *arXiv preprint arXiv:2403.15444*, 2024.

[14] Fatemeh Shafizadegan, Ahmad R Naghsh-Nilchi, and Elham Shabaninia. Multimodal vision-based human action recognition using deep learning: a review. *Artificial Intelligence Review*, 57(7):178, 2024.

[15] Zehua Sun, Qiuhong Ke, Hossein Rahmani, Mohammed Bennamoun, Gang Wang, and Jun Liu. Human action recognition from various data modalities: A review. *IEEE transactions on pattern analysis and machine intelligence*, 45(3):3200–3225, 2022.

[16] Quan Kong, Ziming Wu, Ziwei Deng, Martin Klinkigt, Bin Tong, and Tomokazu Murakami. Mmact: A large-scale dataset for cross modal human action understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8658–8667, 2019.

[17] Seyedmajid Hosseini, Raju Gottumukkala, Satya Katragadda, Ravi Teja Bhupatiraju, Ziad Ashkar, Christoph W Borst, and Kenneth Cochran. A multimodal sensor dataset for continuous stress detection of nurses in a hospital. *Scientific Data*, 9(1):255, 2022.

[18] Attila Reiss and Didier Stricker. Introducing a new benchmarked dataset for activity monitoring. In *2012 16th international symposium on wearable computers*, pages 108–109. IEEE, 2012.

[19] 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.

[20] Francisco Javier Ordóñez and Daniel Roggen. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1):115, 2016.