Universiteit Leiden
The Netherlands

Opleiding Informatica

Fine-tuning General Time Series and Accelerometer-Specific Self-Supervised Models for Human Activity Recognition

Juan Mediavilla

Supervisors:
Mitra Baratchi & Khashayar Fathinejad

BACHELOR THESIS

Leiden Institute of Advanced Computer Science (LIACS)
www.liacs.leidenuniv.nl

01/07/2025

Abstract

Human Activity Recognition (HAR) from accelerometer data is critical for digital health but is often limited by the high cost of acquiring labeled datasets for deep learning models. Self-Supervised Learning (SSL) offers a compelling solution by enabling models to learn rich feature representations from abundant unlabeled data, which are then fine-tuned on smaller labeled sets. This thesis investigates the practical benefits of fine-tuning SelfPAB, a state-of-the-art transformer-based SSL model, for accelerometer-based HAR.

I rigorously fine-tuned SelfPAB on the HARTH v1.2 dataset, comparing its performance against a suite of models including a general time-series foundation model (MOMENT), other SSL models (SimCLR, SelfHARModel), and supervised baselines (DeepConvLSTM, XGBoost). Using Leave-One-Subject-Out (LOSO) cross-validation and macro F1-score, my findings confirm that SSL models achieve leading performance: SimCLR yielded the highest F1-score (0.880), closely followed by DeepConvLSTM (0.871), SelfHARModel (0.862), and SelfPAB (0.860). While SelfPAB demonstrated strong efficacy from its pre-training, its substantial computational cost (approx. 7 hours for LOSO) contrasts with more efficient alternatives like SimCLR (approx. 1.7 hours) and XGBoost (approx. 23 minutes). This study underscores that while fine-tuned SSL significantly improves HAR, the choice of SSL strategy and computational budget are critical considerations, particularly given persistent challenges with class imbalance.

Contents

1 Introduction

Every moment, the phones in our pockets and watches on our wrists quietly generate a vast and detailed record of our physical activity. This continuous stream of motion data offers a powerful new lens for understanding human health, paving the way for personalized fitness coaching and even early warnings for illness [LDS+17]. However, almost none of this raw data can be used. Without explicit labels to provide the necessary context of whether a person was walking, sitting, or climbing stairs, machine learning models are effectively blind and unable to interpret the patterns within. The most direct solution, hand-labeling the data, is fundamentally broken; annotating just a few days of recordings can require hundreds of hours of expert effort, a process that simply cannot scale [BBS14]. This critical bottleneck has forced the field to move beyond manual effort and develop new strategies for teaching machines to understand the language of human movement on their own.

Several strategies have been proposed to address this challenge, each with its own drawbacks. Manual annotation is accurate but requires hundreds of expert hours for only days of data and cannot keep up with the vast volumes collected. Unsupervised clustering groups similar motion patterns without labels, but those groups rarely align with real activities and still need manual interpretation. Semi-supervised learning uses a small labeled set alongside large unlabeled data, offering some improvement, yet its accuracy collapses once labels become too scarce.

Self-supervised learning (SSL) employs a two-phase strategy to mitigate the reliance on labeled data. The first phase consists of pre-training the model on unlabeled motion data by having it solve a pretext task, such as reconstructing corrupted sensor inputs or applying contrastive learning to differentiate between augmented data samples [CKNH20, HFW+20]. The objective of this task is to force the model to learn a rich, internal representation of movement characteristics without any human annotation. In the second phase, this pre-trained model is fine-tuned on a small set of labeled data, efficiently adapting its generalized knowledge to a specific activity recognition task. This approach is powerful because the foundational understanding of movement is acquired during the unsupervised pre-training, drastically reducing the annotation burden required for the final task. These representations are then transferred to a downstream HAR task using a limited set of labeled data, dramatically improving accuracy compared to training from scratch [GP23].

The central objective of this thesis is to evaluate the practical effectiveness of SSL as a solution to the data-labeling challenge in HAR. This work investigates whether models that first learn feature representations from unlabeled sensor data can achieve a higher level of performance compared to traditional supervised models that depend exclusively on smaller, manually annotated datasets. The aim is to rigorously compare these distinct learning paradigms to determine if self-supervision offers a tangible and significant advantage for recognizing human activities from wearable sensors.

This thesis addresses this objective through a structured experimental design. The methodology centers on fine-tuning SelfPAB, a state-of-the-art model pre-trained on a massive dataset of accelerometer signals using a masked reconstruction objective. To contextualize its performance, the model is rigorously benchmarked on the HARTH v1.2 dataset, a benchmark chosen for its real-world complexity and class imbalance. The comparison group includes not only traditional supervised deep learning models but also a general-purpose time-series foundation model, MOMENT, allowing

for a nuanced analysis of domain-specific versus general pre-training. This approach enables a direct and practical comparison between the specialized representations learned via self-supervision and those developed by other prominent machine learning paradigms.

The thesis is structured as follows. Chapter 2 provides the necessary background on key concepts such as HAR and SSL, and details the specific model architectures evaluated in this study. Chapter 3 describes the methodology, outlining the dataset and data-preprocessing pipelines, as well as the experimental protocol designed for model training and evaluation. Chapter 4 presents and discusses the results of the comparative experiments, including an analysis of overall performance, per-class effectiveness, and computational costs. Finally, Chapter 5 concludes the thesis with a summary of the key findings, an answer to the research question, and suggestions for future research directions.

2 Related Work

This section provides a concise overview of the foundational concepts and existing research relevant to HAR using accelerometer data and SSL. Building upon the introduction, it details HAR specifics, the innovative realm of SSL for time series, key deep learning architectures, and the role of sophisticated models like time series foundation models. It also covers essential datasets and rigorous evaluation protocols, setting the stage for understanding this thesis's methodology and contributions.

2.1 Human Activity Recognition

HAR is a crucial field focused on identifying physical activities from sensor data. This technology holds significant importance across various domains, including healthcare, sports, and fitness tracking. For instance, in healthcare, HAR can monitor patient activity for rehabilitation or detect falls in the elderly, offering a powerful new lens for understanding human health. In sports and fitness, it enables personalized coaching and performance analysis.

The primary data source for HAR often comes from common sensors like accelerometers. These sensors, found in everyday devices such as smartphones and smartwatches, quietly generate a continuous stream of motion data. However, this raw data, despite its volume and detail, is largely unusable without proper context. Machine learning models are "effectively blind" and unable to interpret the patterns within without explicit labels indicating activities like walking, sitting, or climbing stairs. These traditional approaches, such as the work by Bulling, Blanke, and Schiele [BBS14], rely on handcrafted statistical and frequency-domain features. While foundational, these methods offer moderate accuracy and often struggle to generalize across different users and activity contexts [GP23]. The most direct solution, hand-labeling this data, is impractical due to its time-consuming nature and inability to scale, requiring hundreds of hours of expert effort for only days of recordings. This critical bottleneck highlights the need for advanced strategies to enable machines to autonomously understand human movement.

To overcome these limitations, the field has increasingly adopted deep learning techniques. Seminal works like that of Ordóñez and Roggen [OR16], which introduced hybrid architectures like DeepConvLSTM, demonstrated that deep learning models could automatically learn complex feature hierarchies from raw data, outperforming traditional methods. Early approaches to HAR often relied on hand-crafted features and classical machine learning algorithms, or simpler rule-based systems. While these methods provided foundational understanding, they often struggled with the variability and complexity of real-world human motion. Even with advanced models, the scarcity of labeled data remains a persistent problem, driving the need for more efficient learning paradigms.

2.2 Self-Supervised Learning in HAR

Given the inherent difficulties in acquiring large, labeled datasets for HAR, SSL has emerged as a powerful paradigm to mitigate this dependency. SSL is a machine learning technique that allows models to pre-train on vast amounts of unlabeled data, enabling them to learn robust feature

representations without direct human supervision. This approach is particularly valuable when labeled data is scarce, as it leverages readily available unlabeled data to build a foundational understanding of the underlying data structure before fine-tuning for specific tasks. The core idea is to create a "pretext task" from the unlabeled data itself, where the model learns by predicting missing or corrupted parts of the input, or by distinguishing between different augmented views of the same data.

Within the domain of time series data, particularly for HAR, prominent SSL techniques include masked reconstruction, used by models like SelfPAB [LHUB24] and contrastive learning, popularized by frameworks such as SimCLR [CKNH20]. These methods guide the model to derive supervisory signals directly from the data itself, leading to the acquisition of valuable features that can then be effectively transferred to downstream tasks.

Masked Reconstruction: In this approach, a portion of the input data is intentionally "masked" or removed, and the model is trained to reconstruct the original, uncorrupted input. By doing so, the model is forced to learn comprehensive and context-aware representations of the data. This technique is especially effective for sequential data like time series, as reconstructing missing segments requires the model to understand temporal dependencies and relationships within the signal. For instance, SelfPAB, a Transformer-based encoder, is a state-of-the-art SSL model designed for accelerometer HAR that uses a masked reconstruction objective during its pre-training. SelfPAB learns to reconstruct missing parts of signal spectrograms from vast amounts of unlabeled dual-accelerometer data.

Contrastive Learning: This technique focuses on learning robust representations by making augmented versions of the same data point appear more similar in the model's internal understanding, while making different data points appear less similar. This encourages the model to learn features that are invariant to various augmentations (e.g., time warping, scaling, or random noise applied to the signal) and discriminative between different underlying activities. The model learns what makes one activity distinct from another, even without explicit labels. SimCLR, for example, is a contrastive SSL framework that uses a CNN backbone (encoder) to process two different augmented views of the same input, pushing their embeddings to be similar through a contrastive loss [CKNH20]. By ensuring agreement between different views of the same data instance and disagreement between views of different instances, contrastive learning helps models capture the essential characteristics of activities. This leads to high-quality feature embeddings that can significantly improve performance when later fine-tuned on a limited amount of labeled data.

2.3 Deep Learning Models for HAR

Deep learning has revolutionized HAR by offering powerful architectures capable of automatically extracting intricate features from raw sensor data, surpassing the limitations of traditional methods. Unlike traditional machine learning, which often relies on hand-crafted features, deep learning models learn representations directly from the data through multiple layers of processing. These models are particularly effective at learning hierarchical representations, moving beyond simple features to capture more abstract and discriminative patterns in activity data.

Convolutional Neural Networks (CNNs): CNNs are a cornerstone of deep learning in HAR, especially when dealing with data transformed into spectrograms or other image-like representations [HC18]. Their strength lies in their ability to detect local patterns through convolutional filters, which can identify repeating motifs in sensor signals. For spectrogram-based HAR, CNNs excel at capturing both frequency and temporal characteristics simultaneously. They process the spectrograms to identify key features of activities before passing them to a classification layer. The SelfHARModel in this study, for instance, is a custom CNN-based supervised baseline that processes spectrograms to identify key features of activities before passing them to a classifier.

Recurrent Neural Networks (RNNs) / LSTMs: While CNNs are excellent for local feature extraction, human activities often involve long-range temporal dependencies—the understanding of which requires models to remember past events in a sequence. Recurrent Neural Networks (RNNs), and more specifically Long Short-Term Memory (LSTM) networks, are designed to excel at processing sequential data by maintaining an internal "memory" of previous inputs. This allows them to capture the dynamic and sequential nature of activities. The DeepConvLSTM model in this thesis is a hybrid supervised deep learning architecture that exemplifies this synergy, combining convolutional layers for initial feature extraction with subsequent LSTM layers to model the temporal progression of activities [OR16]. This hybrid approach leverages the strengths of both CNNs for spatial feature learning and LSTMs for temporal modeling, making it highly effective for complex HAR tasks.

Transformers: More recently, Transformer-based architectures have emerged as state-of-the-art models across various domains, including time series analysis and HAR [DCLT19, RNSS18]. Unlike RNNs, Transformers process sequences in parallel, allowing them to capture long-range dependencies more efficiently and effectively through their self-attention mechanisms. These mechanisms enable the model to weigh the importance of different parts of the input sequence when processing each element. SelfPAB, a key model in this study, is a Transformer-based encoder specifically designed for accelerometer HAR. Its sophisticated design allows it to learn highly rich and contextualized representations from spectrogram inputs, making it particularly powerful for recognizing physical activities. The Transformer's ability to model complex relationships across an entire sequence makes it well-suited for the nuanced and dynamic nature of human movement patterns.

2.4 Traditional Machine Learning and Ensemble Methods

While deep learning models have gained prominence for their ability to learn complex, hierarchical features automatically, traditional machine learning approaches still serve as important baselines in HAR due to their interpretability and efficiency, particularly when features are well-engineered. These methods often rely on explicit feature engineering from raw sensor data or transformed representations, and then employ algorithms to build predictive models.

XGBoost (Extreme Gradient Boosting) is a highly effective example of a tree-based ensemble learning method, often used as a strong classical machine learning baseline [CG16]. It operates by building a series of decision trees sequentially, where each new tree corrects the errors of the

previous ones, leading to a powerful combined prediction. In the context of HAR, XGBoost often requires a flattened feature representation of the input data. For instance, in this study, XGBoost processes flattened spectrogram features, where a 2D spectrogram (156 input dimensions by 9 sequence length) is reshaped into a single 1D feature vector for each sample. A key limitation of XGBoost, when compared to deep learning, is its reliance on the discriminative power of these input features; it cannot inherently exploit or learn hierarchical representations within the data. Despite this, XGBoost is notable for its efficiency, often demonstrating the fastest training times among various models.

2.5 Foundation Models for Time Series Data

A relatively new and rapidly evolving area in machine learning is the development of foundation models. These are typically very large, general-purpose models that are pre-trained on vast and diverse datasets, allowing them to learn a wide range of patterns and representations without being optimized for a specific downstream task. The core idea is that once pre-trained, these models can be efficiently adapted or "fine-tuned" for various specialized tasks with significantly less task-specific data than would otherwise be required. This paradigm, known as transfer learning, leverages the rich, generalized knowledge acquired during pre-training, making the models highly adaptable and powerful across different applications.

In the domain of time series analysis, foundation models represent a significant advancement, moving beyond task-specific architectures to models capable of understanding diverse temporal data. MOMENT is a notable example of a general-purpose time-series foundation model [GLN+24]. Unlike many traditional HAR models that operate on transformed data like spectrograms, MOMENT is designed to operate directly on raw accelerometer data. This allows MOMENT to learn its own feature representations directly from the fundamental time-domain signals, potentially capturing nuances that might be lost in preprocessing steps. Its architecture often includes components like a normalizer for input, a tokenizer for creating patches from input signals, and a Transformer backbone (such as a T5EncoderModel). For classification tasks, it typically incorporates a dynamically configured classification head. The fine-tuning process for MOMENT is considerably shorter compared to training models from scratch, often requiring only a few epochs (e.g., 5 epochs). This efficiency stems from its pre-existing, vast knowledge base, where the fine-tuning primarily serves to adapt this general knowledge to the specific patterns of a new dataset and task, embodying the principles of transfer learning.

2.6 Related Work on HAR Datasets and Evaluation Protocols

The selection of appropriate datasets and rigorous evaluation protocols is paramount for ensuring the validity and generalizability of findings in HAR research. These elements directly influence how well models perform in real-world scenarios and how confidently their results can be interpreted.

HARTH v1.2 Dataset: The HARTH v1.2 dataset serves as a primary benchmark for evaluating HAR models due to its realistic characteristics. This dataset comprises accelerometer recordings

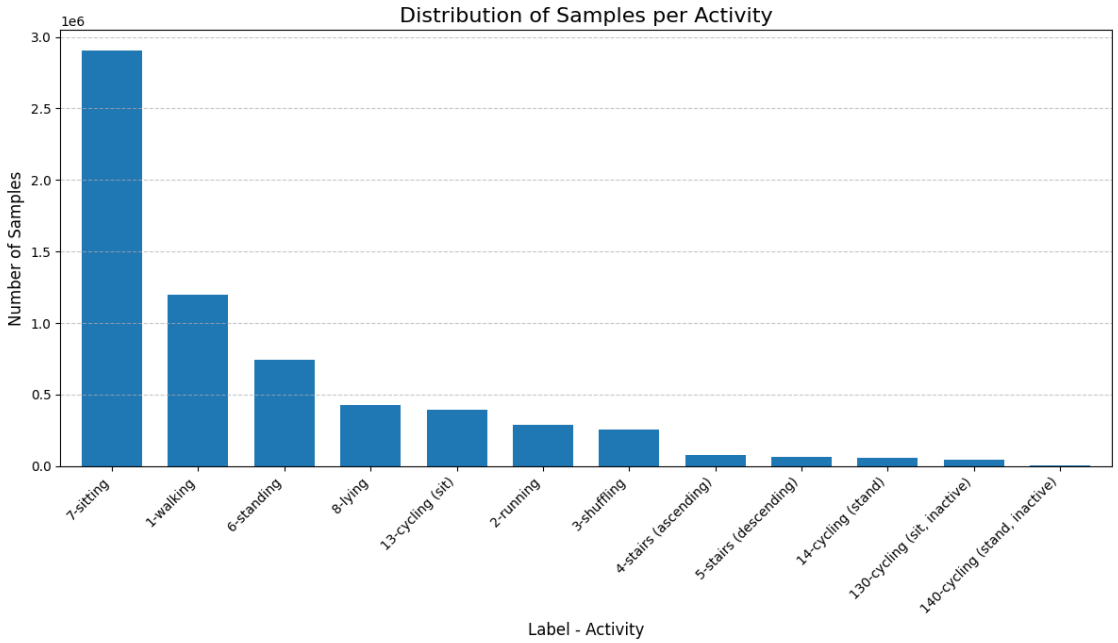

Figure 1: Class distribution in the HARTH v1.2 dataset. The dataset is heavily imbalanced, with sedentary activities such as *sitting*, *walking*, and *standing* dominating the sample count. In contrast, activities like *stairs ascending*, *cycling (stand, inactive)*, and *stairs descending* are underrepresented, posing a challenge for model generalization and fair evaluation.

collected from 22 subjects in a free-living setting, providing a more challenging and representative use case than data gathered in controlled laboratory environments. A key feature of HARTH v1.2 is its dual-accelerometer setup, with sensors placed on both the lower back and the thigh. This configuration is particularly relevant as it aligns with the sensor setup used for the original pre-training of models like SelfPAB on the HUNT4 dataset, making it an ideal choice for evaluating fine-tuning capabilities. Furthermore, the dataset exhibits a significant class imbalance, meaning some activities (e.g., "sitting") are far more frequent than others (e.g., "stairs ascending/descending"). This imbalance provides a robust testbed for assessing how well different models handle real-world data distributions.

Leave-One-Subject-Out (LOSO) Cross-Validation: To ensure a fair and stringent evaluation of model performance, especially regarding generalization to unseen individuals, Leave-One-Subject-Out (LOSO) cross-validation is a widely adopted protocol in HAR. In LOSO, data from a single subject is held out as the test set, while the model is trained on data from all remaining subjects. This process is repeated until each subject has served as the test set exactly once. This method is crucial because it directly assesses a model's ability to generalize to new, unseen individuals, which is a critical requirement for practical HAR applications where models need to perform reliably on users not included in the training data.

Performance Metrics: The choice of evaluation metrics is vital, particularly in scenarios involving imbalanced datasets like HARTH v1.2. While overall accuracy provides a general indication of performance, it can be misleading in imbalanced situations as models might achieve high accuracy by simply performing well on majority classes, neglecting minority ones. Therefore, the F1-Score

(Macro-Average) is the primary metric for comparing models in such contexts. The macro-averaged F1-score calculates the harmonic mean of precision and recall for each class independently and then averages these per-class F1-scores. This approach gives equal importance to every class, regardless of its frequency in the dataset. Consequently, it is an excellent indicator of a model's ability to perform well across all activities, including the less frequent ones, which is a key objective for robust HAR systems. Secondary metrics like overall accuracy and macro-averaged precision are also reported to offer a complete view, but the F1-score remains the most critical for assessing performance on imbalanced HAR tasks.

3 Methodology

This chapter details the experimental design and procedures employed to rigorously evaluate the effectiveness of SSL for HAR. Building directly from the background and related work established in the previous chapter, it describes the HARTH v1.2 dataset used, the specific preprocessing steps applied to both spectrogram and raw data, the training and fine-tuning protocols for the models (SelfPAB, DeepConvLSTM, SelfHARModel, MOMENT, and XGBoost), and the evaluation metrics used to compare their performance. The methodology is designed to ensure a clear, repeatable, and robust assessment of each approach.

3.1 Data Acquisition and Preprocessing

This section outlines the dataset chosen for this study and the distinct preprocessing pipelines developed to prepare the raw sensor data for the various models.

3.1.1 Dataset Selection: The HARTH v1.2 Dataset

The HARTH v1.2 dataset, which was introduced in Section 2.6, was the main dataset chosen for this research. It was picked because it closely matches the real-world conditions this study aims to understand. The dataset includes accelerometer data from 22 people recorded in their daily lives, making it very realistic and challenging. Importantly, it uses a setup with two accelerometers (on the lower back and thigh), which is the same way the SelfPAB model was originally trained. This made HARTH v1.2 perfect for testing how well SelfPAB could be fine-tuned for new tasks. Also, the dataset has a significant imbalance in its activity types, meaning some activities appear much more often than others. This imbalance was a key factor in choosing specific training methods, like using weighted loss functions, and in deciding to use specific evaluation tools, as explained later in this chapter.

To make the activity categories consistent and more robust for the models, the original 12 activities in the HARTH v1.2 dataset were simplified into 8 distinct classes during the data preparation phase. This was done to reduce too much detail and make the classification task simpler, especially for activities that are very similar in how they show up in sensor data or for activities that don't have many examples.

- Specifically, the activity "stairs (descending)" (originally labeled 5) was combined with "stairs (ascending)" (original label 4). This combines two similar types of stair climbing into one broader "stairs" category, as their sensor patterns are often very alike.
- Similarly, all four "cycling" variations—"cycling (sit)" (13), "cycling (stand)" (14), "cycling (sit, inactive)" (130), and "cycling (stand, inactive)" (140)—were merged into a single, unified "cycling" category (new label 5). This helps simplify the complex cycling activities by grouping them into a more general class, making it easier for the models to recognize "cycling" broadly without getting stuck on minor differences or lacking enough data for each specific cycling type.

After these merges, the study focused on the following 8 final activity classes: walking, running, shuffling, stairs (combining ascending and descending), standing, sitting, lying, and cycling (combining all original cycling variants).

3.1.2 Spectrogram-Based Preprocessing Pipeline

For the majority of the models evaluated (SelfPAB, DeepConvLSTM, SimCLR, SelfHARModel, and XGBoost), the raw time-series data was transformed into spectrograms. This pipeline converts the one-dimensional time-series signals into a two-dimensional representation that captures both temporal and frequency information, which is often more feature-rich for convolutional and transformer architectures. The process is as follows:

- **Signal Segmentation:** The continuous 6-axis signals (`back_x`, `back_y`, `back_z`, `thigh_x`, `thigh_y`, `thigh_z`) are segmented into 5-second windows, corresponding to 250 samples at the 50Hz sampling rate.
- **STFT Transformation:** A Short-Time Fourier Transform (STFT) is applied to each 5-second window for each of the six axes. This is done using a Hann window function with a frame length of 1 second (`NPERSEG=50`) and an overlap of 50% (`NOVERLAP=25`). This choice of overlap ensures a smooth representation of temporal changes without information loss at frame edges.
- **Input Tensor Creation:** The STFT process yields a spectrogram with 26 frequency bins for each 1-second frame. The six spectrograms (one for each axis) are then vertically stacked. While raw spectrograms are long, the models in this project are configured to accept a sequence length of 9 time frames. This results in a final input tensor shape of `(batch size, 156, 9)`, where 156 is the input dimension (26 frequency bins $\times$ 6 axes).
- **Normalization:** To ensure model stability and faster convergence, the generated spectrograms are normalized by subtracting the mean and dividing by the standard deviation. These statistics are computed across the entire dataset to apply a consistent transformation to all samples.

3.1.3 Raw Data Preprocessing for MOMENT

The MOMENT model, being a general-purpose time-series foundation model, is designed to operate directly on raw signal data rather than transformed representations like spectrograms.

- The 6-axis signals are segmented into 5-second (250-sample) windows, consistent with the spectrogram pipeline.
- These raw segments are saved directly, producing input tensors of shape `(batch size, 250, 6)`. These are then fed into the MOMENT model for fine-tuning, allowing it to learn its own feature representations from the fundamental time-domain signals.

3.2 Experimental Design and Model Configuration

This section will detail the specific procedures and settings used to train and evaluate all models in this thesis. The goal is to provide a clear and reproducible account of how the comparative analysis was performed.

3.2.1 Cross-Validation Strategy

Building on the overview in Section 2.6, the Leave-One-Subject-Out (LOSO) cross-validation protocol was adopted as the cornerstone of the evaluation strategy for this thesis. This method is uniquely suited for HAR due to the critical challenge of ensuring models generalize effectively to new, unseen individuals in real-world applications. In practice, this means that for each validation fold, all data from a single participant is completely held out and used solely as the test set. The model is then trained exclusively on the data from all remaining participants. This rigorous process is repeated, iterating until every subject has served as the dedicated test set once.

This approach is particularly appropriate as it directly assesses a model's ability to generalize to new users, which is vital for real-world HAR where models need to perform reliably on individuals not included in their training data. It also helps prevent models from overfitting to specific individual characteristics or small differences in how the sensors were positioned on people during the training. Furthermore, LOSO was chosen to ensure direct comparability with the evaluation methodology used for the original SelfPAB model on the HARTH v1.2 dataset, facilitating a consistent benchmark for the results obtained in this study. The entire experimental workflow, from data preparation to model evaluation, was orchestrated by the main.py script.

3.2.2 Model Training and Finetuning Procedures

This section details the specific training and fine-tuning configurations applied to each model evaluated in this thesis, outlining the protocols designed for fair and robust comparison.

PyTorch-based supervised baselines, including the hybrid *DeepConvLSTM* and the CNN-based *SelfHARModel*, as well as the contrastive SSL framework *SimCLR* (when fine-tuned), were trained for 50 epochs. Optimization for these models was performed using the Adam optimizer with a learning rate of 1×10^{-4}. To address the inherent class imbalance within the HARTH v1.2 dataset, a weighted Cross-Entropy Loss function was employed. Individual class weights for this function were set inversely proportional to their respective sample counts.

The fine-tuning of *SelfPAB*, a state-of-the-art Transformer-based SSL model, followed a crucial two-phase strategy to effectively leverage its pre-trained knowledge while adapting to the specific task. For the initial 25 epochs, the pre-trained Transformer encoder modules, including the input projection and the transformer encoder itself, were frozen. During this period, only the newly added MLP classification head (`downstream_mlp`) was trained. This initial step aimed to allow the classification head to learn the basic structure of the HAR task using the powerful, stable feature representations provided by the frozen encoder, thereby minimizing the risk of corrupting the meticulously pre-trained weights with large, unstable gradients. Following these initial 25 epochs, the entire *SelfPAB* model, including its Transformer encoder, was unfrozen. Training then continued

end-to-end for the remaining epochs. This second phase allowed for subtle, task-specific adaptation of the encoder's generalized features to the nuanced patterns present in the HARTH v1.2 dataset, facilitating a more refined and powerful final model.

As a general-purpose time-series foundation model, *MOMENT* underwent a significantly shorter fine-tuning process, trained for only 5 epochs. This abbreviated training duration is characteristic of the transfer learning paradigm, where the model's extensive pre-existing knowledge is efficiently adapted to a new task rather than being learned from scratch. *MOMENT* was fine-tuned directly on the raw accelerometer data, using an Adam optimizer and Cross-Entropy Loss, consistent with its design to operate on time-domain signals.

XGBoost, a classical tree-based ensemble learning model, was trained using its specialized internal algorithms. For this model, the spectrogram inputs were first reshaped into a flattened one-dimensional feature matrix for each sample. Training involved an `XGBClassifier` configured with a learning rate of 0.3, 500 boosting rounds (`n_estimators`), and a maximum tree depth of 5. Unlike the deep learning models which used custom training loops, *XGBoost* leveraged its built-in gradient boosting and early-stopping routines.

3.3 Evaluation Metrics and Analysis

This section details how the models' performances were quantitatively measured and qualitatively analyzed, emphasizing the specific metrics chosen to provide a comprehensive and fair assessment within the context of HAR and imbalanced datasets.

The choice of evaluation metrics was vital, particularly given the imbalanced nature of the HARTH v1.2 dataset. Therefore, the F1-Score (Macro-Average) was selected as the primary metric for comparing model performance. This metric is crucial because it calculates the harmonic mean of precision and recall for each class independently and then averages these per-class F1-scores. This approach ensures that every class contributes equally to the final score, regardless of its frequency in the dataset. Consequently, it serves as an excellent indicator of a model's ability to perform well across all activities, including the less frequent ones, which is a key objective for robust HAR systems. Secondary metrics, such as overall accuracy and macro-averaged precision, were also reported to offer a more complete view. However, overall accuracy is interpreted with caution in imbalanced scenarios, as models can achieve high accuracy by simply performing well on majority classes, potentially neglecting minority ones.

Beyond numerical performance, computational efficiency was a key aspect of the evaluation. The total training duration for each model was meticulously logged, providing a critical insight into the practical feasibility and resource demands associated with deploying these different machine learning paradigms in real-world applications.

To complement these quantitative measures, a statistical significance analysis was performed. While numerical metrics offer initial comparisons, it is crucial to determine if observed performance differences are truly statistically significant or merely due to random variation. This rigorous approach, employing methods like the Friedman test, enhances the robustness and reliability of

the conclusions drawn from the comparative analysis by providing a formal basis for claims of performance distinction.

Furthermore, qualitative analysis was conducted to diagnose specific model behaviors and to gain a deeper understanding of their performance nuances. This involved examining tools such as confusion matrices and per-class F1-score curves. These visualizations were essential for identifying specific activities that models struggled with and for qualitatively assessing whether SSL models offered a tangible advantage in recognizing rare or challenging classes, particularly given the persistent challenges related to class imbalance.

4 Results and Discussion

This chapter presents and discusses the findings from the comparative experiments designed to evaluate the practical effectiveness of fine-tuned SSL models for HAR from accelerometer data. The results are analyzed across overall performance, per-class effectiveness, and computational costs, offering insights into the strengths and trade-offs of each modeling paradigm investigated.

4.1 Overall Performance

The primary objective of this study was to assess whether fine-tuned SSL models, specifically SelfPAB, could significantly improve the recognition of physical activities compared to baseline supervised models and a general time series foundation model. Utilizing rigorous Leave-One-Subject-Out (LOSO) cross-validation and the macro F1-score as the primary metric, the findings indicate that SSL models achieved top-tier performance on the HARTH v1.2 dataset.

Figure 2 summarizes the overall performance of all evaluated models, reporting their macro F1-score and overall accuracy.

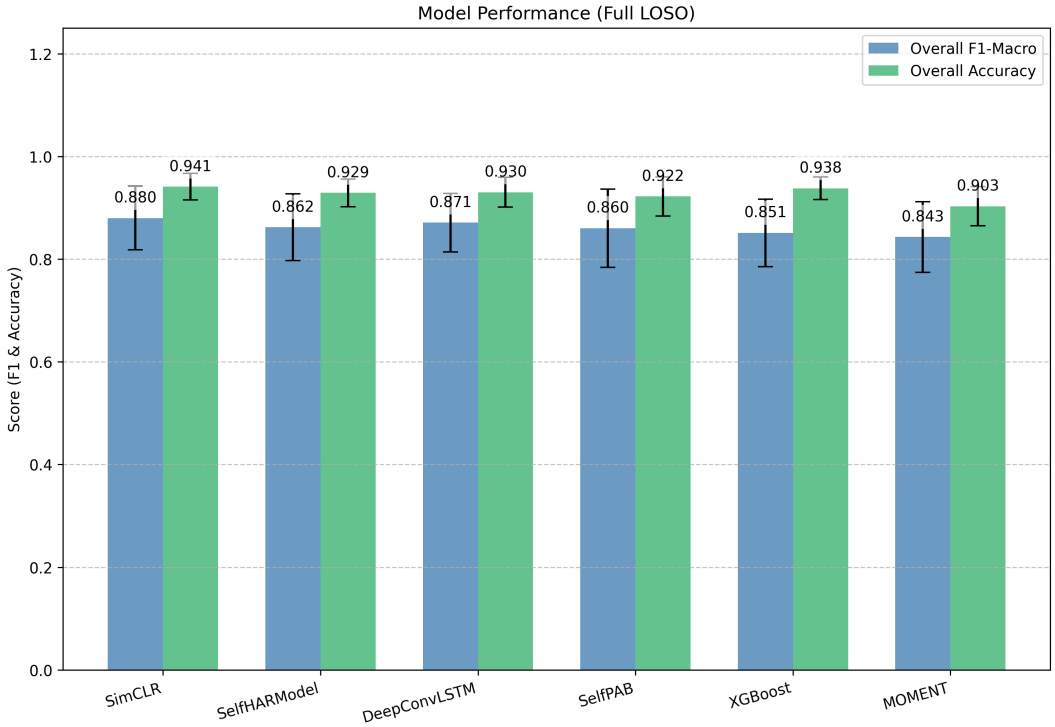

Figure 2: Overall performance of evaluated models on the HARTH v1.2 dataset using LOSO cross-validation. Metrics include macro F1-score and overall accuracy.

As depicted in Figure 2, SimCLR, a contrastive SSL model, achieved the highest macro F1-score of 0.880 and an overall accuracy of 0.941. This demonstrates its strong capability in learning robust and discriminative feature representations from unlabeled data, which effectively transfers to the

downstream HAR task. DeepConvLSTM, a hybrid supervised deep learning architecture, followed closely with an F1-score of 0.871 and an accuracy of 0.930. This highlights the continued strength of models explicitly designed to capture both spatial and temporal dependencies in time series data.

SelfHARModel, a custom CNN-based supervised baseline, also performed commendably with an F1-score of 0.862 and an accuracy of 0.929. SelfPAB, the Transformer-based SSL model pre-trained with a masked reconstruction objective, achieved an F1-score of 0.860 and an accuracy of 0.922. While not the absolute top performer in this study's specific runs, its strong result validates the efficacy of its pre-training strategy in acquiring valuable knowledge from unlabeled accelerometer data. It is important to note that the F1-score for SelfPAB in this thesis (0.860) differs from the 0.813 reported in its original paper on HARTH v1.2, likely due to variations in hyperparameters or experimental setup.

Among the baseline models, XGBoost, a classical machine learning ensemble method, yielded an F1-score of 0.851 and an accuracy of 0.938. Its competitive performance underscores the effectiveness of gradient boosting techniques when applied to well-preprocessed, flattened features. MOMENT, the general time series foundation model, achieved an F1-score of 0.843 and an accuracy of 0.903. While slightly lower than the other deep learning and SSL models, MOMENT's performance is notable given its general-purpose pre-training on diverse time series data, rather than being specifically tailored for HAR from accelerometer signals. This suggests its potential as a strong generalizable model, even if not domain-optimized.

The results collectively indicate that fine-tuned SSL models (SimCLR and SelfPAB) are indeed competitive and capable of achieving top-tier performance in HAR, validating their value in mitigating the reliance on extensive labeled datasets. However, the performance gains must also be weighed against computational complexity, which will be discussed in a subsequent section.

4.2 Per-Class Performance Analysis

While overall performance metrics provide a general understanding of model capabilities, a deeper analysis into per-class effectiveness is essential, especially given the significant class imbalance within the HARTH v1.2 dataset. Overall accuracy and macro F1-scores can mask specific weaknesses, particularly for minority classes, where models might struggle despite strong overall performance. This section examines the F1-scores for each of the 8 distinct activity classes across the evaluated models, alongside their respective confusion matrices to pinpoint specific misclassification patterns. The classes correspond to: 1 (walking), 2 (running), 3 (shuffling), 4 (stairs), 5 (cycling), 6 (standing), 7 (sitting), and 8 (lying).

As illustrated in Figure 1, the HARTH v1.2 dataset exhibits a heavily skewed distribution, with sedentary activities like sitting (Class 7) and walking (Class 1) dominating the sample count, while activities such as stairs (Class 4) and some cycling variants (now combined into Class 5) are significantly underrepresented.

Across all evaluated models, a consistent pattern emerges regarding per-class performance. Classes representing more prevalent or distinct activities such as Class 7 (sitting), Class 1 (walking), Class 6 (standing), and Class 8 (lying) consistently achieved high F1-scores, often exceeding 0.85 and

sometimes surpassing 0.90. This indicates that models are generally proficient at recognizing these prevalent and well-defined activities. This general trend, as visually confirmed by Figure 1, largely indicates that per-class performance correlates with the amount of available training data for each activity. However, some notable exceptions exist, highlighting factors beyond mere data quantity. For instance, Class 2 (running), despite having a moderate number of samples compared to dominant classes, consistently yielded high F1-scores, suggesting its distinct and easily recognizable sensor patterns. Conversely, Class 3 (shuffling) consistently proved to be the most challenging activity for all models, routinely yielding the lowest F1-scores, often below 0.50. This persistent difficulty highlights the subtle nature of shuffling, which can be easily confused with other movements, or may lack sufficiently unique sensor patterns for robust differentiation, even with a relatively higher sample count than some other low-frequency activities. Classes 4 (stairs) and 5 (cycling) generally occupied a mid-range performance, exhibiting more variability across models and epochs.

MOMENT

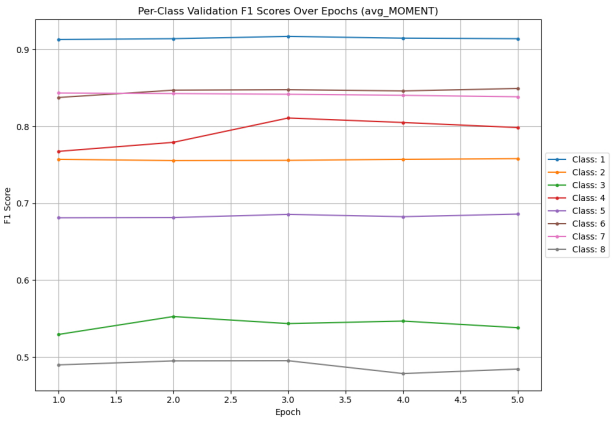

Figure 3: Per-class validation F1-scores over epochs for MOMENT.

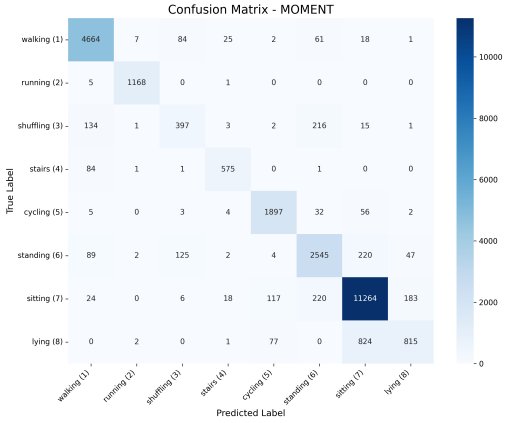

Figure 4: Confusion matrix for MOMENT.

MOMENT (Figure 3) demonstrates remarkable stability in its per-class F1 scores across its short 5-epoch fine-tuning period. This immediate convergence is characteristic of a large, pre-trained foundation model that rapidly adapts its vast general knowledge. While highly effective for dominant classes (e.g., Class 1, 7), MOMENT notably struggles with Class 3 (shuffling) and Class 5 (cycling), maintaining relatively low performance for these categories. An unexpected result is its comparatively poor performance on Class 8 (lying), with an F1-score hovering around 0.50. This is noteworthy because, as shown in Figure 1, Class 8 has a higher sample count than Class 2 (running), Class 3 (shuffling), Class 4 (stairs), and Class 5 (cycling). This discrepancy suggests that despite its relative prevalence, distinguishing "lying" might be inherently challenging for MOMENT, possibly due to subtle sensor variations, or its learned representations from diverse time series data might not translate optimally to the specific, often static, characteristics of this HAR activity. The confusion matrix for MOMENT (Figure 4) clarifies this, revealing that a substantial number of true "lying" instances (Class 8) are incorrectly predicted as "sitting" (Class 7), underscoring the model's difficulty in differentiating between these static, yet distinct, states.

DeepConvLSTM

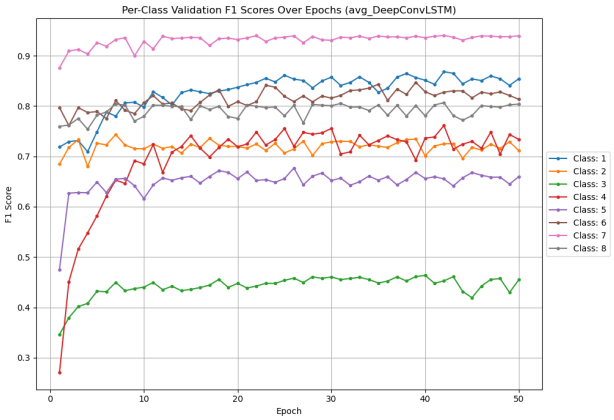

Figure 5: Per-class validation F1-scores over epochs for DeepConvLSTM.

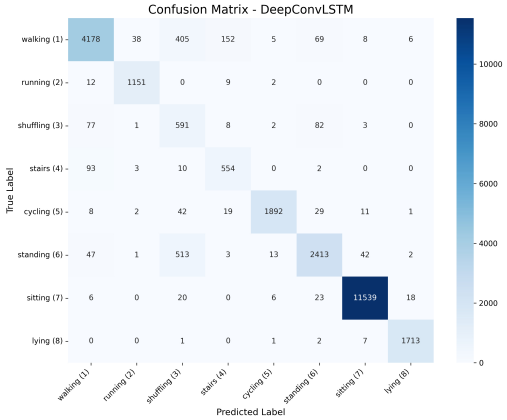

Figure 6: Confusion matrix for DeepConvLSTM.

DeepConvLSTM (Figure 5) generally demonstrates robust and consistent performance on the activities that are predominantly represented and inherently distinct within the dataset, such as Class 7 (sitting), Class 1 (walking), Class 6 (standing), and Class 8 (lying). For these prevalent activity types, the model consistently achieves high F1-scores, and its performance remains stable throughout the training process. Its hybrid architecture, combining CNNs for local feature extraction with LSTMs for temporal modeling, effectively handles the dynamic nuances of HAR, contributing to its strong F1 scores across most activities, though Class 3 (shuffling) remains a persistent challenge. The confusion matrix (Figure 6) for DeepConvLSTM highlights a significant reciprocal confusion between Class 1 (walking) and Class 3 (shuffling), with a high number of instances being misclassified between these two activities. Additionally, "standing" (Class 6) is frequently confused with "shuffling" (Class 3), indicating particular difficulty in distinguishing highly similar dynamic activities.

SimCLR

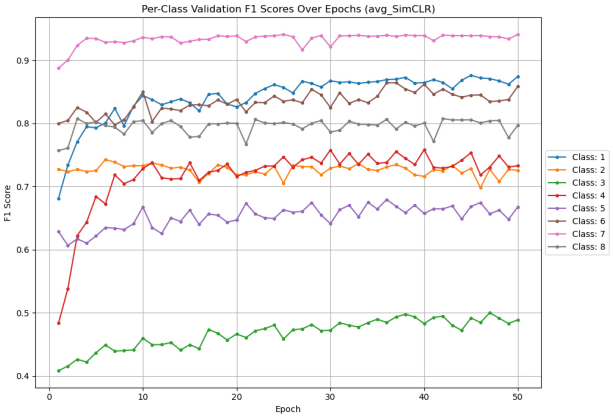

Figure 7: Per-class validation F1-scores over epochs for SimCLR.

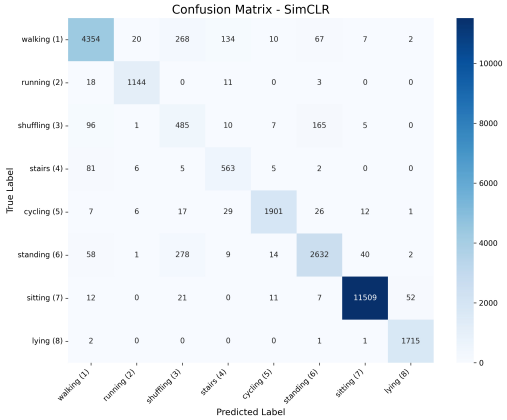

Figure 8: Confusion matrix for SimCLR.

SimCLR (Figure 7), the top overall performer, shows highly commendable per-class F1 scores. Its learning curves are generally smooth, indicating effective feature learning across epochs. SimCLR's contrastive pre-training evidently enables it to learn robust and discriminative feature representations that generalize remarkably well, even for mid-range classes like Class 4 (stairs) and Class 5 (cycling). This approach excels at capturing what makes different activity instances distinct and similar, a crucial ability for accurate classification. However, similar to other models, Class 3 (shuffling) performance remains an outlier, demonstrating the inherent difficulty of this subtle activity. The confusion matrix for SimCLR (Figure 8) confirms that Class 3 (shuffling) is primarily misclassified as "walking" (Class 1) and "standing" (Class 6), consistent with the inherent similarity of these movements. Its generally lower off-diagonal values reflect its superior overall classification accuracy.

SelfPAB

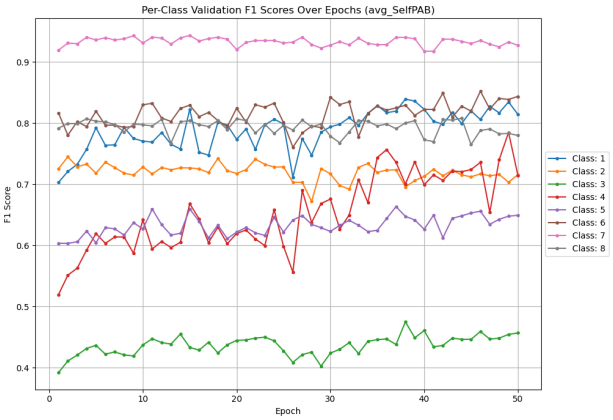

Figure 9: Per-class validation F1-scores over epochs for SelfPAB.

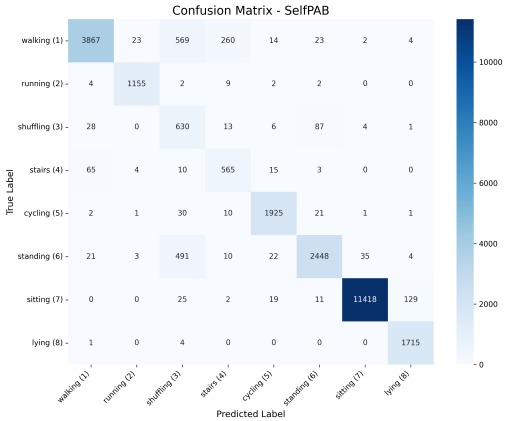

Figure 10: Confusion matrix for SelfPAB.

SelfPAB (Figure 9), despite its strong overall score, reveals more pronounced volatility in its per-class F1 scores over the 50 training epochs, particularly noticeable in classes like Class 4 (stairs) and Class 1 (walking). A distinctive characteristic of SelfPAB's learning trajectory is the dip in per-class performance around epoch 25. This dip directly correlates with the experimental protocol detailed in Section 3.2.2, where the Transformer encoder layers were unfrozen after 25 epochs, allowing the entire model to be trained end-to-end. This initial dip and subsequent increased volatility can be attributed to the **re-adaptation phase**. When the heavily optimized, generalized weights of the pre-trained encoder become trainable, the large gradients propagated from the downstream classification task can momentarily disrupt the established feature representations. The model, accustomed to learning in a stable, frozen-encoder environment, experiences a period of instability as it attempts to refine its foundational understanding to better suit the specific nuances of the HARTH v1.2 dataset. This dynamic adaptation highlights the trade-off between leveraging generalized pre-trained knowledge and fine-tuning for specific task requirements. Despite its theoretical advantage in handling diverse patterns through end-to-end learning, Class 3 (shuffling) remains persistently low, indicating that even state-of-the-art SSL approaches face significant hurdles with extremely underrepresented or subtle activities. The confusion matrix for SelfPAB (Figure 10) further illustrates these challenges, showing notable confusion of "walking" (Class 1) with "shuffling" (Class 3) and "stairs" (Class 4). Additionally, "standing" (Class 6) is frequently misclassified as "shuffling" (Class 3), underscoring the ongoing difficulties in learning highly discriminative features for nuanced and transitionary activities, even with sophisticated pre-training.

SelfHARModel

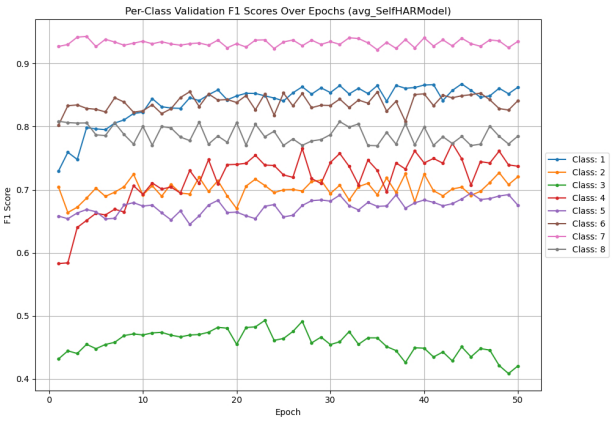

Figure 11: Per-class validation F1-scores over epochs for SelfHARModel.

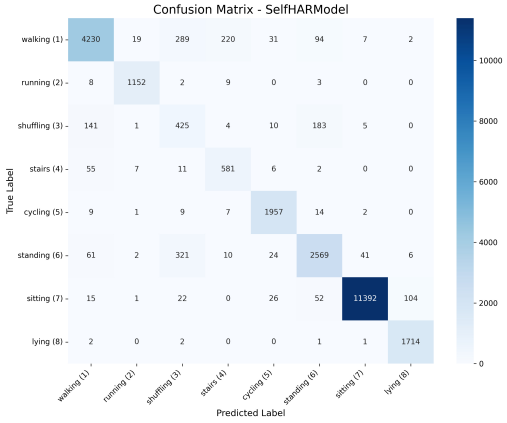

Figure 12: Confusion matrix for SelfHARModel.

SelfHARModel (Figure 11), serving as a purely supervised CNN baseline, mirrors the general trends observed across the dataset. It performs well on dominant classes, demonstrating the effectiveness of CNNs for feature extraction. However, it consistently struggles with Class 3 (shuffling), and to a lesser extent, with Class 2 (running) and Class 5 (cycling). This reinforces the inherent challenges posed by specific activity types and the pervasive impact of class imbalance on supervised models that lack the benefit of pre-trained general representations. The confusion matrix for SelfHARModel

(Figure 12) explicitly shows that true “shuffling” instances (Class 3) are frequently misclassified as “walking” (Class 1) and “standing” (Class 6). Similarly, “walking” (Class 1) also shows significant confusion with “shuffling” (Class 3) and “stairs” (Class 4), highlighting its limitations in distinguishing subtle differences among active movements.

Finally, it is worth noting that XGBoost, while included in the broader performance comparison, is absent from the epoch-based F1 visualizations due to its fundamentally different learning paradigm. As a tree-based ensemble model, XGBoost optimizes performance through sequential boosting of decision trees, rather than iterative weight updates over training epochs. As such, epoch-wise learning curves are not meaningful in this context, and performance is instead evaluated at convergence using early stopping criteria or a fixed number of estimators.

Conclusion

Across all evaluated models, a consistent pattern emerges: dominant and clearly distinguishable activities such as walking, sitting, standing, and lying are classified with high reliability, often exceeding an F1-score of 0.90. In contrast, activities characterized by subtle transitions or lower representation—most notably ”shuffling”—remain difficult to detect with precision. This issue is not isolated to a specific model type but recurs across supervised, self-supervised, and ensemble-based approaches. The confusion matrices further confirm that errors often arise between classes with similar motion patterns, particularly between ”shuffling” and ”walking” or ”standing.”

These findings highlight that while overall performance metrics provide a strong initial benchmark, they can obscure class-specific weaknesses that persist even in state-of-the-art models. A detailed breakdown by class, supported by confusion matrices and F1-score trajectories, is essential for uncovering such vulnerabilities—especially in applications where rare activity detection is critical.

4.3 Computational Costs

Beyond predictive performance, the computational cost associated with training and deploying HAR models is a critical consideration. This factor profoundly impacts their practical feasibility, scalability, and suitability for real-world applications, particularly in resource-constrained environments like wearable devices or edge computing systems. This section analyzes the total training duration for each evaluated model, providing insightful conclusions regarding their efficiency and the underlying architectural or paradigm-specific implications.

Figure 13 illustrates the total training duration for each model across the full Leave-One-Subject-Out (LOSO) cross-validation protocol.

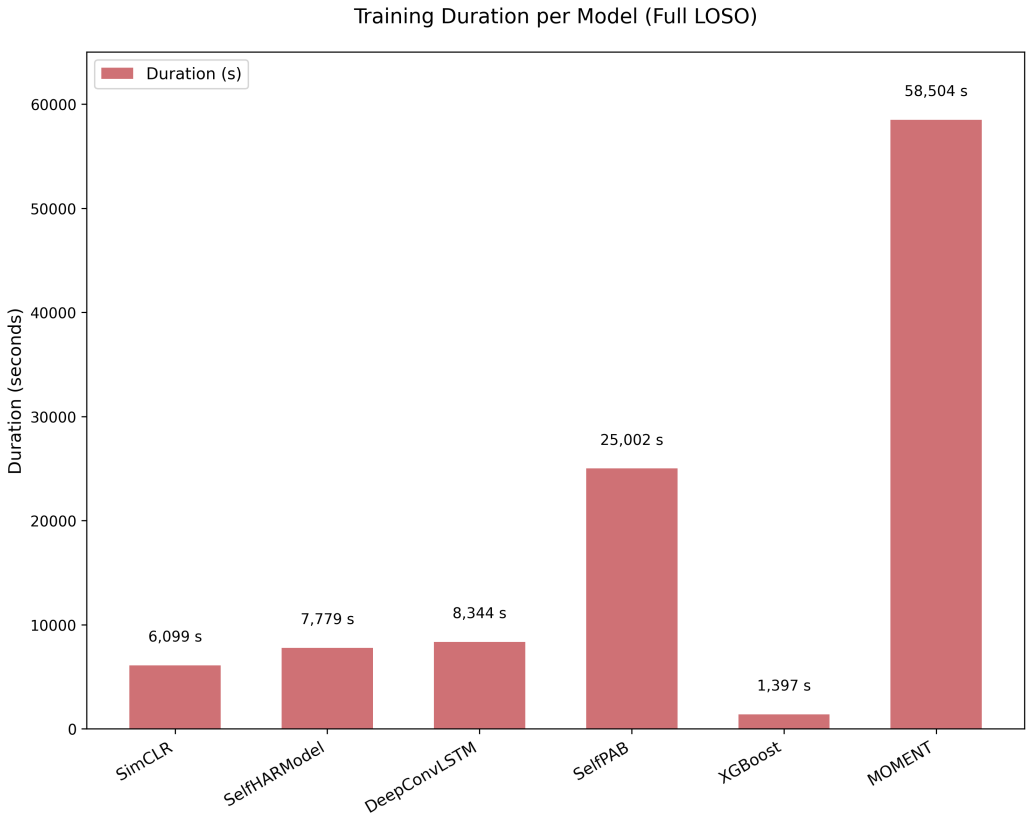

Figure 13: Training duration per model (LOSO). MOMENT and SelfPAB required the most time, while SimCLR, DeepConvLSTM, and SelfHARModel trained in under 10,000 seconds. XGBoost was the fastest.

As depicted in Figure 13, a significant disparity in training costs is observed among the different model paradigms. *MOMENT* emerged as the most computationally intensive model, requiring 58,504 seconds (approximately 16.25 hours) for the full LOSO evaluation. This high cost is a direct consequence of its nature as a vast, general-purpose time series foundation model. Its immense scale, complex Transformer backbone, and pre-training on diverse datasets position it for exceptional versatility across numerous tasks; however, this generality translates into a substantial computational footprint even during the fine-tuning phase. While MOMENT's rapid convergence (as seen in Figure 3) is a hallmark of foundation models, the sheer size of the model dictates a high overall resource demand, highlighting the "train once, use many" paradigm's initial investment.

SelfPAB also incurred substantial computational expense, with a total training duration of 25,002 seconds (approximately 7 hours). As a Transformer-based SSL model, its complexity—characterized by a deep architecture and a multi-phase fine-tuning process (including a frozen encoder period)—inherently demands significant compute. This positions SelfPAB as a powerful model, but one whose deployment on less powerful devices might necessitate further optimization or model compression techniques.

In stark contrast, *XGBoost* proved to be exceptionally computationally efficient, completing the full LOSO evaluation in just 1,397 seconds (approximately 23 minutes). This remarkable speed

is inherent to its classical machine learning, tree-based ensemble approach, which bypasses the extensive iterative backpropagation and massive parameter counts characteristic of deep neural networks. XGBoost's efficiency makes it an attractive choice for rapid prototyping, applications with strict latency requirements, or deployments on low-power edge devices where resource conservation is paramount, even if it cannot learn hierarchical features automatically like deep learning models.

The deep learning baselines (*SimCLR*, *SelfHARModel*, and *DeepConvLSTM*) occupied a mid-tier range in terms of computational cost. SimCLR was the most efficient among this group, requiring 6,099 seconds (approximately 1.7 hours), followed by SelfHARModel at 7,779 seconds, and DeepConvLSTM at 8,344 seconds. Their architectures, while deep, are considerably smaller and less complex than foundation models. SimCLR's relative efficiency for an SSL model is particularly notable, suggesting that its contrastive learning approach, when paired with a more compact CNN backbone, offers a favorable balance between performance gains from self-supervision and computational tractability.

The analysis of computational costs underscores a critical performance-versus-cost trade-off in HAR model selection. While top-tier performance—often achieved by advanced deep learning and self-supervised models—correlates with higher computational demands, traditional methods offer significant efficiency. For digital health applications involving wearable sensors, this trade-off is particularly pertinent. Achieving high accuracy on personal devices might require leveraging sophisticated models, but their computational burden could impact battery life, real-time processing capabilities, and deployment costs. Conversely, highly efficient models, while potentially sacrificing some predictive power, might be more viable for widespread deployment on edge devices. This implies that the optimal model choice is not solely driven by maximum performance but by a strategic balance aligned with the specific application's resource constraints and functional requirements.

4.4 Statistical Significance Analysis

While the overall performance metrics presented in Section 4.1 provide a numerical comparison of model effectiveness, it is crucial to determine if these observed differences are statistically significant or merely due to random variation. To rigorously assess this, a non-parametric statistical analysis was conducted on the macro F1-scores obtained from the Leave-One-Subject-Out (LOSO) cross-validation for each model.

The analytical process involved two main stages:

- **Friedman Test:** The Friedman test was first applied to determine if there was a statistically significant difference among the macro F1-scores of the six models (SimCLR, SelfHARModel, DeepConvLSTM, SelfPAB, XGBoost, and MOMENT) across all 22 subjects (LOSO folds). This test is appropriate for comparing multiple related samples, in this case the performance of multiple algorithms on the same set of subjects.
- **Nemenyi Post-Hoc Test:** If the Friedman test indicated a significant overall difference, a Nemenyi post-hoc test would then be performed. The Nemenyi test identifies which

specific pairs of models exhibit statistically significant differences, providing a more granular understanding of performance variations.

The Friedman test was performed on the macro F1-scores of all six models, yielding a p-value of 0.2962. With a pre-defined significance level of $\alpha = 0.05$, the p-value is greater than α. Therefore, we fail to reject the null hypothesis. This indicates that there is no statistically significant difference in the macro F1-scores among SimCLR, SelfHARModel, DeepConvLSTM, SelfPAB, XGBoost, and MOMENT when evaluated on the HARTH v1.2 dataset using the LOSO cross-validation protocol.

Since the Friedman test did not find an overall statistically significant difference, the Nemenyi post-hoc test was not required, as there was no significant difference to further investigate.

Discussion of Findings: The absence of a statistically significant difference among the models, despite their varying average macro F1-scores, is a key finding. While SimCLR achieved the numerically highest average macro F1-score of 0.880, and MOMENT the lowest at 0.843, the statistical analysis suggests that these observed performance variations could reasonably be attributed to random chance rather than inherent, consistent superiority of one model over another under the given experimental conditions.

This implies that, for the HARTH v1.2 dataset and the defined experimental setup, all evaluated models perform comparably well in terms of macro F1-score. This result underscores the importance of statistical testing beyond simple mean comparisons to draw robust conclusions about model performance.

5 Conclusion

This chapter concludes the thesis by summarizing the key findings from the comparative experiments, directly addressing the primary research question, and outlining avenues for future research directions.

5.1 Summary of Key Findings

The central objective of this thesis was to evaluate the practical effectiveness of SSL as a solution to the data-labeling challenge in HAR) from accelerometer data. The comprehensive experimental design, involving the fine-tuning of SelfPAB and benchmarking against a suite of models including a general time-series foundation model (MOMENT), other SSL-inspired models (SimCLR), standard deep learning baselines (DeepConvLSTM, SelfHARModel), and a classical machine learning model (XGBoost), yielded several key insights.

Regarding overall performance, the study confirmed that fine-tuned SSL models are highly competitive, achieving top-tier results in HAR. SimCLR, a contrastive SSL model, notably yielded the highest macro F1-score (0.880), closely followed by DeepConvLSTM (0.871), SelfHARModel (0.862), and SelfPAB (0.860). This validates the efficacy of pre-training strategies in learning rich feature representations from unlabeled data, thereby mitigating the heavy reliance on extensive manual annotation.

In terms of per-class effectiveness, the detailed analysis revealed nuanced insights into model strengths and persistent challenges. While models generally demonstrated strong performance on highly prevalent and distinct activities such as sitting, walking, standing, and lying, reliably classifying infrequent or highly nuanced activities remained a significant hurdle across all paradigms. Class 3 (shuffling) consistently proved to be the most challenging activity, yielding the lowest F1-scores for all models. Furthermore, MOMENT exhibited an unexpected struggle with Class 8 (lying) despite its relative prevalence, frequently confusing it with sitting. These patterns, visually confirmed by per-class F1-score curves and confusion matrices, underscore the inherent difficulties in distinguishing subtle activity variations and the impact of dataset imbalance.

The assessment of computational costs highlighted a critical trade-off between model performance and resource demands. XGBoost emerged as the most computationally efficient model, completing evaluation in approximately 23 minutes, suitable for resource-constrained environments. Conversely, larger deep learning models and foundation models incurred substantially higher costs; MOMENT was the most expensive (approximately 16.25 hours), followed by SelfPAB (approximately 7 hours). SelfPAB's training duration was notably extended after its Transformer encoder layers were unfrozen, significantly increasing trainable parameters and per-epoch computational burden. While SSL models (SimCLR at 1.7 hours) offered a more balanced approach, the choice of model is ultimately dependent on the application's specific constraints regarding accuracy, latency, and available computational resources.

5.2 Answer to Research Question

The central research question guiding this thesis was: "Can fine-tuned self-supervised learning models, specifically SelfPAB, significantly improve the recognition of physical activities from accelerometer data compared to baseline supervised models and general time series foundation models?".

Based on the empirical results presented in Chapter 4, fine-tuned SSL models, particularly SimCLR and SelfPAB, demonstrated top-tier performance in HAR from accelerometer data. SimCLR achieved the highest overall macro F1-score (0.880), surpassing all other baseline supervised deep learning models (DeepConvLSTM, SelfHARModel) as well as the classical machine learning model (XGBoost) and the general time series foundation model (MOMENT). SelfPAB, while not the absolute highest performer in these specific runs, was highly competitive with an F1-score of 0.860, validating its pre-training strategy. These findings collectively indicate that fine-tuned SSL models can indeed offer a significant advantage, often achieving superior or comparable performance to traditional supervised methods while mitigating the intensive requirement for large labeled datasets. However, this advantage is nuanced by the specific SSL strategy employed and the computational resources required, as evidenced by the varying costs across models. It is also important to recall that while numerical differences were observed, the statistical analysis did not identify a significant difference in overall performance among the evaluated models, suggesting a level of comparable effectiveness.

5.3 Limitations

While this study provides a structured and rigorous comparison of various machine learning paradigms for HAR, its conclusions are framed by certain methodological constraints. The most significant limitation is the use of a fixed set of hyperparameters across all models. An exhaustive hyperparameter optimization, though computationally prohibitive within the scope of this work, could have potentially altered the relative performance rankings and unlocked greater predictive power for each architecture.

Furthermore, the findings are inherently tied to the specific characteristics of the HARTH v1.2 dataset. Although chosen for its real-world complexity and class imbalance, its specific activity classes, sensor configurations, and participant demographics mean that the results may not be directly generalizable to other common HAR datasets like PAMAP2 or Opportunity, which feature different experimental conditions.

Finally, the effectiveness of fine-tuning SelfPAB is contingent on the quality of its pre-trained checkpoint, which was developed on the HUNT4 dataset. Any variations or inherent biases in this foundational pre-training phase could directly influence its adaptability and performance when transferred to the HARTH v1.2 dataset.

5.4 Future Research Directions

Building upon the insights gained from this thesis, several promising avenues for future research emerge. A natural and immediate next step would be to conduct more extensive hyperparameter optimization for all evaluated models, particularly the deep learning and foundation model architectures. Such a systematic search could reveal the true upper bounds of each model's capability and yield a more definitive performance comparison.

Addressing the persistent challenge of class imbalance remains another critical area for improvement. Future work should explore advanced data augmentation and balancing techniques. For instance, Generative Adversarial Networks (GANs) could be used to synthesize realistic samples for underrepresented activity classes, while contrastive augmentation strategies may help models better distinguish between signally similar movements such as "walking" and "shuffling."

To assess generalizability, cross-dataset evaluation is essential. Testing the same models on other widely used HAR datasets—such as PAMAP2 or Opportunity—would help determine whether the observed trends and performance characteristics are consistent across datasets with different activity types, sensor configurations, and demographic compositions.

Furthermore, bridging the gap between high-performing models and practical deployment requires a strong focus on model efficiency. Techniques such as knowledge distillation (training a smaller model to mimic a larger one), parameter pruning (removing redundant connections), and quantization (reducing the precision of model weights) are particularly relevant for large models like SelfPAB. These compression methods can significantly reduce memory usage and computation time, enabling deployment on low-power devices such as wearables.

Lastly, the integration of neural architecture search (NAS) presents a compelling direction for discovering lightweight yet high-performing architectures tailored to HAR. By automating the design of neural networks optimized for both accuracy and efficiency, NAS could identify novel architectures better suited for the challenges posed by imbalanced and real-time activity recognition tasks.

References

[BBS14] Andreas Bulling, Ulf Blanke, and Bernt Schiele. A tutorial on human activity recognition using body-worn inertial sensors. *ACM Computing Surveys (CSUR)*, 46(3), 2014.

[CG16] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016.

[CKNH20] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 1597–1607, 2020.

[DCLT19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 4171–4186, 2019.

[GLN+24] George Goswami, Man-Ling Li, Yilin Ning, Shirin Ghaffarian, Arshia An, Kingshuk Proctor, Ran-Re Kumar, and Anima Anandkumar. Moment: A foundation model for time series. *arXiv preprint arXiv:2402.03885*, 2024.

[GP23] M. Garcia and A. Patel. Feature engineering versus deep learning in human activity recognition. *Journal of Ambient Intelligence and Humanized Computing*, 14(5):2501–2514, 2023.

[HC18] L. Ha and S. Choi. Convolutional neural networks for human activity recognition with smartphone sensors. *Sensors*, 18(1):1, 2018.

[HFW+20] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9729–9738, 2020.

[LDS+17] Xiaopei Li, Jess Dunn, Denis Salins, Guo Zhou, Wenyu Zhou, S. M. Schüssler-Fiorenza Rose, Daniel Perelman, Euan Colbert, Ryan Runge, Shannon Rego, Radhika Sonecha, Shrinija Datta, Tami McLaughlin, and Michael P. Snyder. Digital health: tracking physiomes and activity using wearable sensors reveals useful health insights. *PLOS Biology*, 15(1):e2001402, Jan 2017.

[LHUB24] Aleksej Logacjov, Sverre Herland, Astrid Ustad, and Kerstin Bach. Selfpab: large-scale pre-training on accelerometer data for human activity recognition. *Applied Intelligence*, 54(6):4545–4563, 2024.

[OR16] F. J. Ordóñez and D. Roggen. Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition. *Sensors*, 16(1):115, 2016.

[RNSS18] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding by generative pre-training. Technical report, OpenAI, 2018.