Human Activity Recognition in Aging Populations via Time-Frequency Domain Features and Ensemble Learning

Jaykumar Darji (018180851)[a], Jainil Rao (018236582)[a]

[a]*Department of Electrical Engineering, San Jose State University, San Jose, CA*

Abstract

Human Activity Recognition (HAR) is essential in healthcare, elder care, and wearable systems to monitor mobility, detect falls, and support activity-based interventions. Age-related changes in mobility patterns and the prevalence of assistive device usage among older adults increase the complexity of accurate activity recognition. This study presents a machine learning framework for activity classification using statistical and frequency-domain features derived from tri-axial accelerometer signals. The HAR70+ dataset, comprising sensor data from the lower back and right thigh of 18 older adult participants (aged 70–95), was used for model development. Signals were segmented into 1-second windows, and 36 time-frequency features were extracted per window, including dominant frequency, spectral energy, mean, standard deviation, median, and mode. Activities were classified into three setups: a fine-grained 7-class model, a simplified 4-class setup, and a binary static vs dynamic classification. Multiple classifiers were evaluated, with Random Forest achieving the highest accuracy and selected as the final model. Hyperparameter tuning was performed using RandomizedSearchCV with 500 iterations, evaluating 7,500 models in total. The top 10 most important features, as determined by Random Forest feature importance, were retained for the final evaluation. Evaluation on the test set produced 95.0% accuracy for the 7-class task, 97.48% for the 4-class task, and 98% for the binary classification task. External validation on a subset of the HARTH dataset yielded 88.89% accuracy, confirming generalization across settings. Results support the effectiveness of time-frequency feature engineering and classical machine learning for real-time, interpretable activity monitoring in aging populations.

Keywords: Human Activity Recognition, Accelerometer Data, Wearable Sensors, HAR70+, Time-Frequency Features, Random Forest, Ensemble Classifier.

1. Introduction

The global adoption of wearable technologies has transformed how physical activity is monitored, particularly in healthcare, rehabilitation, and elderly support systems [1, 2, 3]. Accelerometer-based activity tracking plays a crucial role in these domains by enabling continuous, non-invasive monitoring of daily movement patterns. Among older adults, such monitoring can assist in fall prevention, functional assessment, and the early detection of mobility-related decline, thus promoting independence and improving care outcomes.

Email addresses: `jaykumar.darji@sjsu.edu` (Jaykumar Darji (018180851)), `jainil.rao@sjsu.edu` (Jainil Rao (018236582))

With advancements in sensor design and placement, body-worn devices now provide high-resolution, multichannel data that can capture subtle differences in posture and mobility. However, translating raw sensor readings into meaningful activity labels remains a complex challenge, especially when accounting for physical variability across age groups, the use of assistive devices, and the natural heterogeneity of movement in free-living conditions. While many existing systems rely on predefined thresholds or cut-points to infer activity levels, such heuristics often underrepresent the physical effort required by older individuals and fail to capture transitional or atypical behaviors.

Machine learning (ML) techniques offer a promising alternative, capable of learning complex relationships between sensor data and activity categories. Most state-of-the-art approaches leverage deep neural networks, which automatically extract discriminative features from raw signals. However, deep learning models typically require large labeled datasets, high computational resources, and may lack interpretability—factors that limit their practical use in real-time and resource-constrained environments such as wearable edge devices.

In contrast, classical ML methods based on handcrafted features provide greater transparency and computational efficiency. Statistical measures (e.g., mean, standard deviation) and frequency-domain descriptors (e.g., dominant frequency, spectral energy) are frequently used due to their simplicity and effectiveness. These features can be extracted from fixed-length signal windows and used to train classifiers with relatively low data and hardware demands. Nevertheless, ensuring generalizability across subjects, especially older adults with varied functional capacities, remains a key concern.

1.1. Literature Review

Human Activity Recognition (HAR) has emerged as a critical area of research due to its potential applications in healthcare, smart environments, education, and security systems. With the proliferation of wearable sensors and smartphones embedded with accelerometers and gyroscopes, the capacity to monitor and classify human actions in real-time has greatly expanded [1, 2, 3]. In healthcare, HAR plays a vital role in patient monitoring and elderly care [4, 5], while also contributing to safe learning environments and infant safety monitoring [6].

Traditional machine learning (ML) techniques have long been used in HAR for effective activity classification. Support Vector Machines (SVMs) remain a popular choice due to their balance between performance and interpretability. For instance, Kusuma et al.[7] enhanced activity classification by tuning SVM parameters, achieving an accuracy of 96.26%. Similarly, Qian et al. [8] introduced a multi-class SVM with a binary tree architecture, improving classification through advanced feature engineering. Efficiency concerns, such as long training times, were addressed by Chathuramali and Rodrigo [9] via feature vector optimization, while Anguita et al. [10] proposed an SVM model specifically optimized for smartphone sensors.

Other ML models have also shown promise in HAR. Random Forests (RF), using bagging and ensemble learning techniques, have proven effective in various studies [11, 12]. Additionally, gradient boosting machines (GBMs) such as LightGBM have been explored for their performance advantages despite high computational costs. Zhang et al.[13] introduced a semi-supervised LightGBM approach to reduce reliance on labeled data. Further enhancements to LightGBM were presented by Shao et al. [14] and Csizmadia et al.[15].

Simpler classifiers, such as Naïve Bayes and K-Nearest Neighbors (KNN), remain relevant due to their interpretability and low computational demand. Maswadi et al. [16]employed Naïve Bayes to predict activities like walking and sitting, while Shen and Fang [17] demonstrated improved results using Gaussian Naïve Bayes. Liu et al. [18] introduced kernel-based enhancements to KNN, and Mohsen et al. [19] applied tuning and evaluation strategies such as ROC analysis to refine KNN performance.

As sensor data has grown in volume and complexity, deep learning (DL) methods have become increasingly dominant in HAR, owing to their capacity to learn intricate data representations. Convolutional Neural Networks (CNNs), capable of capturing spatial dependencies, were effectively used by Ronao and Cho [20] in smartphone-based HAR. Similarly, Ha and Choi [21] demonstrated improved results using CNNs on multi-sensor data. Recurrent Neural Networks (RNNs), particularly LSTM and GRU architectures, are well-suited for modeling sequential patterns, as seen in DeepConvLSTM by Ordóñez and Roggen [22].

To further enhance HAR accuracy, attention-based mechanisms have been integrated with RNNs. Zeng et al. [23] proposed AttentionLSTM, which selectively focused on critical temporal features. Transfer learning also emerged as a strategy to address data scarcity, as illustrated by Morales and Roggen [24], who leveraged pre-trained CNNs. Hybrid models, such as EnsembleLSTM by Guan and Plötz [25], combined the strengths of various architectures to boost recognition performance.

The relevance of DL in elder care has been specifically highlighted. Awais et al. [26] achieved 90% accuracy in classifying older adults' daily activities using SVM and inertial sensors. Cherian et al. [27] investigated daily behaviors in individuals with dementia, noting Random Forest's superiority among ML classifiers. Noury et al. [28] evaluated multiple ADLs, revealing stronger classification performance in tasks like toileting and dressing when using combined sensor data.

Further advances in DL for HAR include models proposed by Mihoub [29] and Gupta [30], who explored CNN-GRU hybrids and feature selection techniques for smart home applications. Dogan et al. [31] stressed the importance of preprocessing in multi-sensor environments. Khan et al. [32] and Durga et al. [33] developed CNN-LSTM and enhanced LSTM models, respectively, using diverse datasets such as UCI-HAR. Rahayu et al. [34] introduced a CNN model focused on body joint features, achieving notable accuracy.

A comparative evaluation by Alam et al. [35] benchmarked traditional ML against DL algorithms, finding that while tree-based methods like C4.5 offered speed and efficiency, neural networks such as ANN and deep ANN (DANN) delivered superior accuracy, albeit with greater computational cost. This trade-off remains a central issue in HAR development.

Another major challenge in both ML and DL methods is the selection of hyperparameters, which significantly influences model performance. Hyperparameter tuning is essential but resource-intensive. Yu and Zhu [36] underscored the unpredictability in DL training and the impact of tuning on accuracy. Cooney et al. [37] demonstrated this in EEG signal classification using CNNs. LiYang et al. [38] introduced Hyper-Tune, a framework aimed at simplifying the tuning process, while Bischl et al. [39] offered a comprehensive analysis of tuning strategies from grid search to advanced methods. Raiaan et al. [40] recently highlighted the specific challenges and progress in hyperparameter tuning for CNNs, indicating its continued importance and complexity in DL applications.

In summary, HAR research is evolving from traditional ML techniques to more advanced DL architectures, driven by the need for greater accuracy, robustness, and adaptability. While DL offers enhanced performance, challenges remain in terms of computational demand, personalization for users (especially older adults), and the efficient tuning of model parameters. Future research should address these limitations to enable scalable, real-time HAR solutions.

This study presents a comprehensive machine learning framework for recognizing human activity from dual accelerometer data collected from the lower back and right thigh of older adults. The proposed approach segments sensor signals using a 1-second sliding window and extracts 36 interpretable time-frequency features per window. Activities are categorized into three levels of complexity: a fine-grained 7-class task, a mid-level 4-class setup, and a binary classification distinguishing between dynamic and static activities.

Multiple classical classifiers are benchmarked, and the best-performing model—Random Forest—is optimized through extensive hyperparameter tuning. The final model incorporates the top 10 most informative features and undergoes further evaluation using an external dataset, HARTH, to test cross-domain generalization. This work demonstrates that carefully engineered classical models can provide accurate, efficient, and deployable solutions for activity monitoring in aging populations, particularly in scenarios where privacy, interpretability, and low computational burden are essential.

2. Materials and Methods used

This section describes the complete methodology used to develop our Human Activity Recognition (HAR) system. It covers the dataset characteristics, data preprocessing pipeline, feature extraction techniques, and the formulation of multiple classification problems. The entire approach is designed around lightweight and interpretable feature engineering, enabling the use of classical machine learning models.

*2.1. **Data Acquisition***

This study is based on the Human Activity Recognition 70+ (HAR70+) dataset [41], a rich and professionally annotated collection of accelerometer recordings from 18 older adults aged 70 to 95. The dataset captures a range of real-world activities performed by participants with varying levels of mobility, including five individuals who used walking aids during the recording sessions. Each participant wore two Axivity AX3 accelerometers, one attached to the lower back and the other to the right thigh. The recordings lasted approximately 40 minutes per subject and were collected during a semi-structured, free-living protocol—encouraging natural movement through a set of predefined activities. Data was sampled at 50 Hz, resulting in a high-resolution time-series suitable for fine-grained motion analysis. To ensure accurate ground truth, sessions were video-recorded using a chest-mounted camera, and the activity labels were assigned frame-by-frame by professional annotators. Each data point includes a timestamp, six acceleration signals (back_x, back_y, back_z, thigh_x, thigh_y, thigh_z), and a corresponding activity label. In total, the dataset comprises 2,259,597 labeled samples [41].

2.2. *Data Preprocessing*

After acquiring and organizing the HAR70+ dataset, we performed several preprocessing steps to prepare the data for machine learning model development. These steps ensured the temporal integrity of the time-series data, removed inconsistencies, and segmented the continuous signal into meaningful activity windows.

2.2.1. *Data Cleaning*

Before segmentation and feature extraction, the raw accelerometer data underwent a thorough cleaning process to ensure data integrity. No missing or null values were detected in any of the sensor recordings across the HAR70+ dataset. Additionally, no duplicate rows were present, confirming the quality and consistency of the collected time-series data.

To assess redundancy and feature relevance, a correlation analysis was conducted between the raw signal axes and the activity labels. Figure 1 shows the correlation matrix computed using Pearson's correlation coefficient. The matrix reveals that two features—`back_y` and `thigh_y`—exhibited negligible correlation with the activity labels and showed high redundancy with other axes. As a result, these two features were excluded from the final analysis to reduce dimensionality and improve model performance.

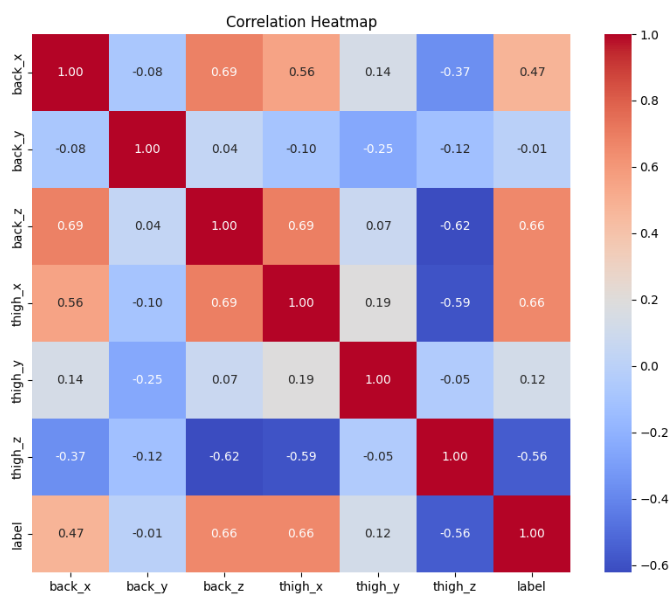


Figure 1: Correlation matrix between raw accelerometer signals and activity labels.

2.2.2. *Subject-wise Splitting*

To avoid data leakage and to ensure generalizability, the dataset was split on a per-subject basis. Specifically, 70% of the subjects were assigned to the training set, and the remaining 30% were reserved for testing.

2.2.3. *Time Sorting and Index Reset*

Each subject's data was sorted chronologically by timestamp before windowing to preserve the natural order of activity transitions. The DataFrame index was reset after sorting to maintain compatibility with downstream processing.

2.2.4. *Sliding Window Segmentation*

The raw time-series data from each subject was segmented using a fixed-length sliding window approach. We experimented with both 1-second and 5-second window sizes (corresponding to 50 and 250 samples, respectively, given the 50 Hz sampling rate). Since both settings yielded comparable results, we opted for the 1-second window to maintain finer temporal resolution. Each window was labeled using majority voting—the most frequent activity label among the samples in the window was assigned as the label for the entire window. This reduced noise in the labels and maintained consistency across samples.

2.2.5. *Short Window Removal*

Windows containing fewer than 10 samples were discarded to ensure that each segment contained sufficient data for meaningful feature extraction. This also avoided instability in calculations such as standard deviation and FFT.

2.3. *Feature Extraction*

After segmenting the raw time-series data into 1-second windows, we extracted a set of features from each segment to represent its underlying motion characteristics. Our goal was to keep the feature set interpretable, compact, and well-suited for machine learning models. We focused on two primary categories of features: statistical features and frequency-domain features, computed independently for each signal axis (back_x, back_y, back_z, thigh_x, thigh_y, thigh_z), as well as for computed magnitude signals (back_mag, thigh_mag).

2.3.1. *Statistical Features*

The following basic statistical descriptors were extracted from each windowed signal:

Let $x = \{x_1, x_2, ..., x_n\}$ be the sequence of values from a single signal axis in a given window (where $n = 50$ samples for a 1-second window at 50 Hz). The following statistical features were extracted:

- **Mean:** Average value of the signal over the window.

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- **Standard Deviation (Std):** Measures variability in the signal.

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$$

- **Median:** Middle value in the ordered list of signal values.

$$\text{Median}(x) = \text{mid}(\text{sort}(x))$$

- **Mode:** Most frequently occurring value in the window.

$$\text{Mode}(x) = \arg\max_{v} \text{count}(x = v)$$

2.3.2. *Frequency-Domain Features*

To capture periodic and oscillatory patterns in the signal, we computed frequency-based features using the Fast Fourier Transform (FFT).

Let $X = \mathrm{FFT}(x)$ be the discrete Fourier transform of x, and let $|X_k|$ denote the magnitude of the k-th frequency component. Frequency-domain features help capture the signal's periodic and dynamic nature:

- **Dominant Frequency:** The frequency component with the highest amplitude (excluding DC).

$$f_{\text{dom}} = f_k \quad \text{where} \quad k = \arg\max_{k>0} |X_k|$$

- **Spectral Energy:** The total signal power in the frequency domain (Sum of squared magnitudes of the FFT components).

$$E = \sum_{k=1}^{n/2} |X_k|^2$$

2.4. *Classification problem*

The activity recognition task in this study was framed as a supervised, multi-class classification problem. Based on the activity labels available in the HAR70+ dataset, we designed and evaluated three different classification settings to examine the model's ability to generalize across granular and abstract activity categories:

- **7-Class Classification:** This setup uses all original annotated activities as distinct classes: walking (1), shuffling (3), stairs ascending (4), stairs descending (5), standing (6), sitting (7), and lying (8).

- **4-Class Classification:** In this setup, related activities were grouped to reduce class overlap and improve robustness. The merged categories were:

    1. Walking-related (includes walking, shuffling, stairs ascending, stairs descending)
    2. Standing
    3. Sitting
    4. Lying

- **Binary Classification:** Finally, a simplified version of the task was created by categorizing all activities into:

    1. Dynamic activities (walking, shuffling, stairs ascending/descending)
    2. Static postures (standing, sitting, lying)

Each of the above classification tasks was treated as an independent problem, with separate training and evaluation performed using the same feature extraction and modeling pipeline. This allowed us to assess the model's performance not only at a fine-grained level but also in broader, more deployment-friendly classification schemes.

2.5. *Model development*

After extracting from the data, We trained and evaluated a variety of classical machine learning models to compare their effectiveness in recognizing human activities across three classification setups: 7-class, 4-class, and binary. The following models were implemented and evaluated: **Random Forest**, **Decision Tree**, **Logistic Regression**, **Support Vector Machine (SVM)** with both linear and RBF kernels, **K-Nearest Neighbors (KNN)**, **Gradient Boosting**, **XGBoost**, and **AdaBoost**. Each model was evaluated using a consistent feature set and data pipeline to ensure fair comparison.

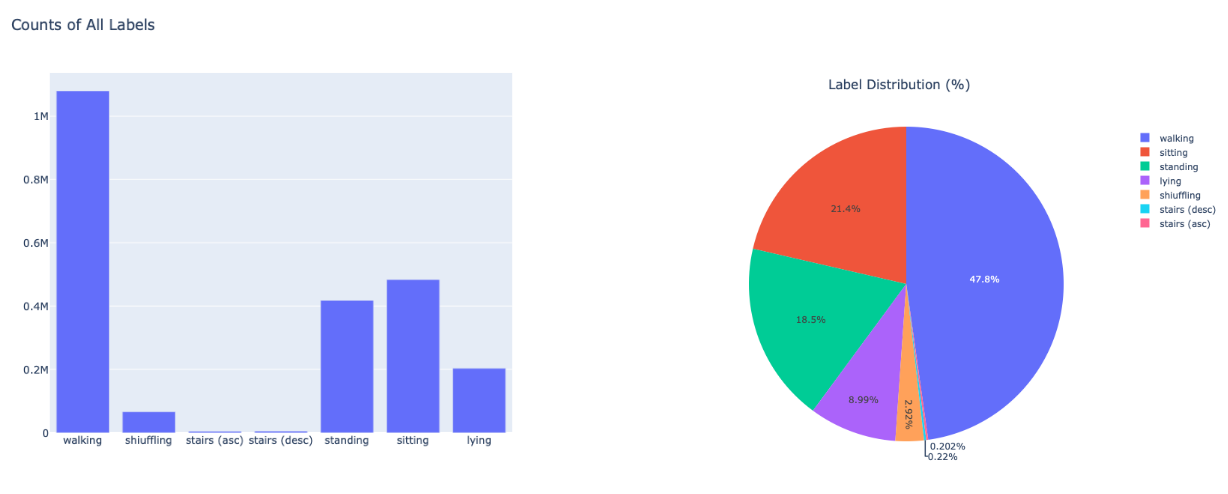
2.5.1. *Data imbalance*

One of the main challenges in training robust HAR models is the presence of class imbalance. In the HAR70+ dataset, dynamic activities such as walking and standing are heavily represented, while minority classes like stair ascending, descending, and shuffling have significantly fewer samples.

We initially experimented with **SMOTE** (Synthetic Minority Over-sampling Technique) to oversample the underrepresented classes. While this approach improved training accuracy, it introduced overfitting during validation and testing, especially for the minority classes. This was likely due to the small number of real samples available to synthesize meaningful variation.

Downsampling the majority classes was also considered, but ultimately avoided to preserve valuable data, as many minority classes already had very few instances.

As a final strategy, we relied on the `class_weight='balanced'` setting in the Random Forest classifier. This approach adjusts the model's internal weighting to penalize misclassification of rare classes more heavily, without modifying the data distribution itself.



(a) Distribution of activity classes in the HAR70+ dataset.

(b) Percentage distribution of activity classes.

Figure 2: Distribution of activity classes in the HAR70+ dataset.

2.5.2. *Cross-Validation Strategy*

To evaluate model performance robustly, we employed **Stratified 5-Fold Cross Validation**. In this approach, the training dataset is split into five equally sized folds, while ensuring that the proportion of each class is preserved within every fold. During each iteration, one fold is used as the validation set, and the remaining four folds are used for training. This process is repeated five times so that every fold is used exactly once for validation.

This strategy is particularly suitable for imbalanced datasets, as it avoids bias introduced by uneven class distributions. The final performance metrics are reported as the average across all five validation runs, providing a stable and generalized estimate of the model's performance.

2.5.3. *Model Selection Criteria*

Among all the models tested, **Random Forest consistently outperformed the others** in terms of both accuracy and generalizability across the 7-class, 4-class, and binary classification problems. Its ability to handle non-linear relationships, feature importance interpretability, and robustness to overfitting made it the preferred model for further optimization and deployment.

2.6. *Hyper parameter tuning*

To further optimize the performance of the selected Random Forest model, we conducted extensive hyperparameter tuning using `RandomizedSearchCV`. Rather than exhaustively searching all combinations in the parameter space, randomized search offers a computationally efficient approach by sampling a fixed number of hyperparameter combinations at random.

The tuning process explored a wide range of hyperparameter values across the following dimensions:

- `n_estimators`: [100, 300, 500, 700, 1000]
- `max_depth`: [10, 20, 30, 40, 50, 60, 70, None]
- `min_samples_split`: [2, 5, 10, 15, 20]
- `min_samples_leaf`: [1, 2, 4, 6, 8]
- `bootstrap`: [True, False]
- `max_features`: ['sqrt', 'log2', 0.6, 0.8, None]
- `criterion`: ['gini', 'entropy']

This grid resulted in over **20,000 possible configurations**. Given the high dimensionality of the search space, we used **RandomizedSearchCV** with **500 random combinations**, each evaluated using 5-fold cross-validation. In total, this required training and evaluating **7,500 models**. The best-performing set of hyperparameters, selected based on cross-validation accuracy, is shown below:

```
RandomForestClassifier(n_estimators=700, min_samples_split=2, min_samples_leaf=1, max_features='sqrt', max_depth=60, criterion='gini', bootstrap=True, class_weight='balanced', random_state=42, n_jobs=-1)
```

The model was then retrained on the full training set using this configuration and evaluated on the held-out test set.

3. Results

This section presents the performance evaluation of various machine learning models trained on the HAR70+ dataset. We report results for three different classification problems: 7-class, 4-class, and binary classification—using statistical and frequency-domain features extracted from 1-second windows of accelerometer data. Each model was evaluated using **Stratified 5-Fold Cross Validation** on the training data to ensure robustness, especially in the presence of class imbalance. The final model, selected based on cross-validation performance, was then retrained on the entire training set and evaluated on a held-out test set to assess generalization. Additionally, we tested the final model on an external dataset, **HARTH**, to evaluate its ability to generalize to unseen data collected in a different setting. This external validation demonstrates the model's potential for real-world deployment.

3.1. Performance of All Models

Table 1 summarizes the test set performance of all evaluated models in terms of overall accuracy, macro-averaged precision, recall, and F1-score. These metrics highlight not just overall correctness, but also how well each model performs across both majority and minority activity classes.

Table 1: Performance metrics for all models on the 7-class classification task (Test Set)

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	0.9054	0.55	0.53	0.51
Gradient Boosting	0.9130	0.60	0.59	0.57
XGBoost	0.9130	0.60	0.59	0.57
Logistic Regression	0.8880	0.50	0.52	0.50
Decision Tree	0.8650	0.56	0.55	0.54
K-Nearest Neighbors	0.8665	0.51	0.52	0.51
SVM (RBF Kernel)	0.8318	0.53	0.63	0.52
SVM (Linear Kernel)	0.7571	0.53	0.61	0.51
AdaBoost	0.8555	0.36	0.40	0.38

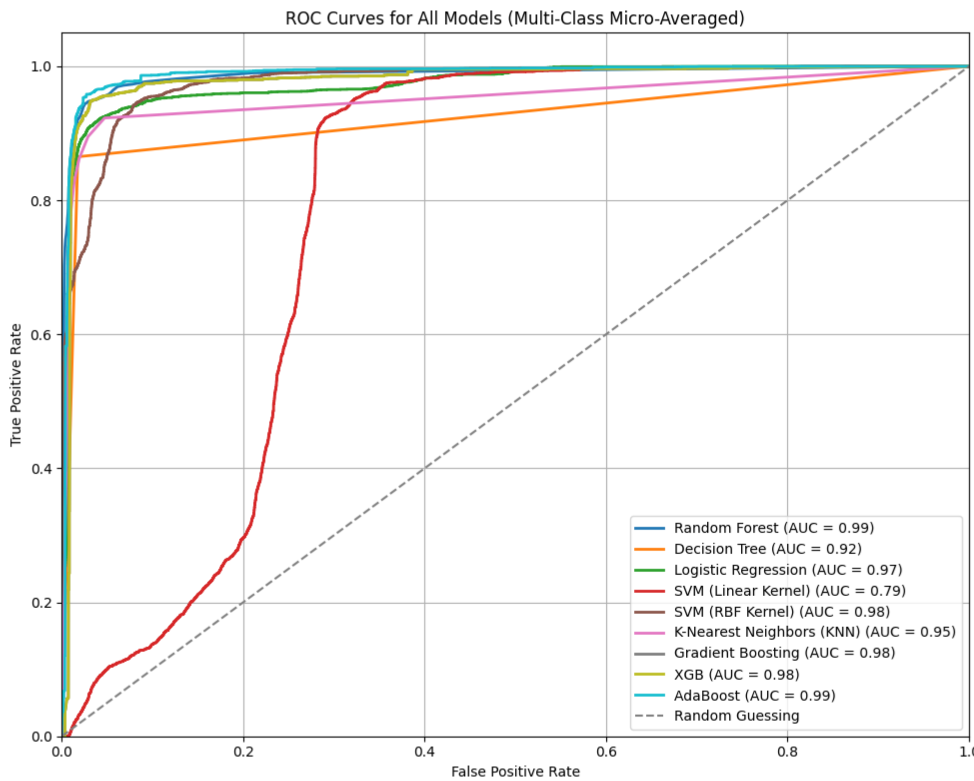


Figure 3: ROC Curve for all the models

As the table 1 indicates, **Gradient Boosting**, **XGBoost**, and **Random Forest** achieved the highest accuracy and macro-averaged scores, making them the most suitable for final deployment. However, all models struggled with minority class recognition, as reflected by relatively low macro recall and F1 scores. Figure 3 displays the ROC curves for all models in the 7-class classification task, providing further insight into their classification boundaries and sensitivity.

Based on the initial evaluation, **Random Forest** and **Gradient Boosting** were selected for hyperparameter tuning, as they consistently outperformed the other models in terms of accuracy and macro-averaged metrics. Both models demonstrated strong generalization capabilities and robustness across all three classification tasks. Following a detailed tuning process using `RandomizedSearchCV`, Random Forest ultimately delivered better overall performance, particularly on the minority classes, and was thus chosen as the final model for deployment and external testing.

3.2. Classification Results After Hyper Parameter Tuning

After hyperparameter tuning, the optimized Random Forest model was evaluated across all three classification setups: 7-class, 4-class, and binary classification. The results reported below are based on Stratified 5-Fold Cross Validation on the training set comprising 29,601 windows extracted from 12 training subjects. Each evaluation used the same feature extraction pipeline consisting of statistical and frequency-domain features.

3.2.1. 7-Class Classification

In the 7-class task (described earlier), the model achieved strong performance on major classes like walking, sitting, and lying. However, precision and recall for underrepresented classes such as stairs ascending, stairs descending, and shuffling remained low due to class imbalance. The model achieved an accuracy of 95.23%, a macro F1-score of 0.68, and a weighted F1-score of 0.94.

3.2.2. 4-Class Classification

In the 4-class setup, walking-related activities were grouped into one class, and the rest were categorized as standing, sitting, or lying. The model demonstrated excellent generalization and consistent performance across all folds. The model achieved an accuracy of 97.48%, a macro F1-score of 0.98, and a weighted F1-score of 0.97.

3.2.3. Binary Classification

For binary classification, activities were grouped into dynamic (walking-related) and static (posture-related) classes. This task yielded the best performance due to simplified class boundaries. The model achieved an accuracy of 97.63%, a macro F1-score of 0.98, and a weighted F1-score of 0.98.

Confusion matrices for all classification setups are shown in Figures 5, 6, and 7, highlighting that most misclassifications in the 7-class task occurred between motion-similar activities, such as walking and stairs. These errors were significantly reduced in the 4-class and binary setups. Table 2 provides class-wise precision, recall, and F1-scores for each classification type. It also compares model performance using all features versus the top 10 most important features (selected using feature importance scores from the Random Forest model).

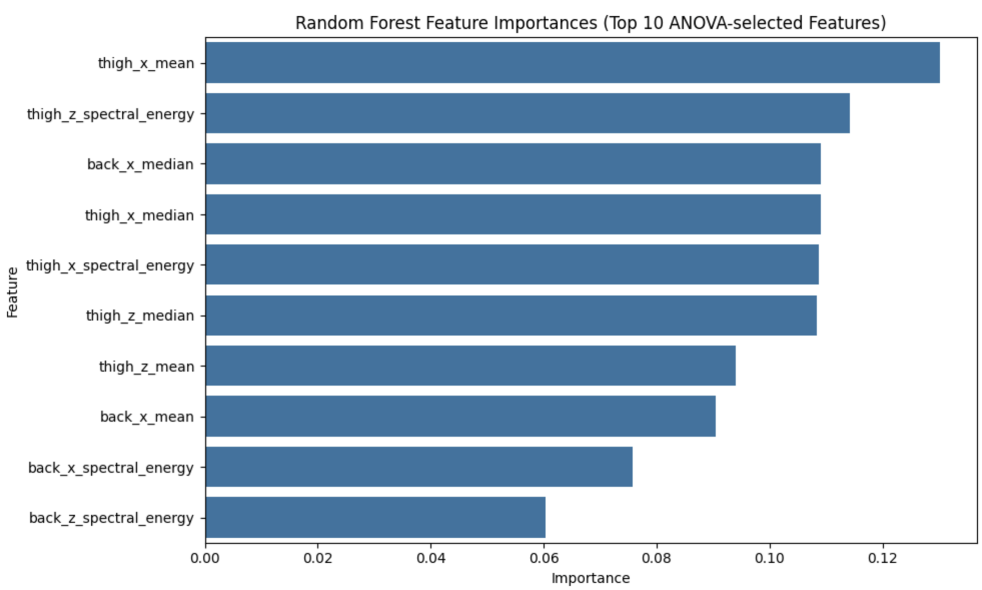
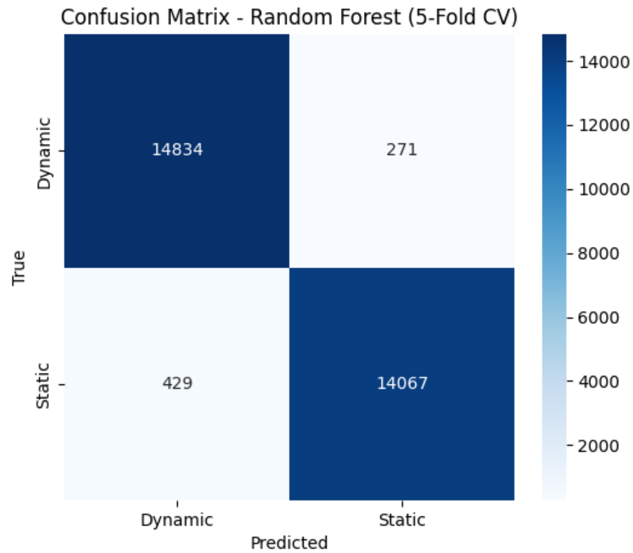


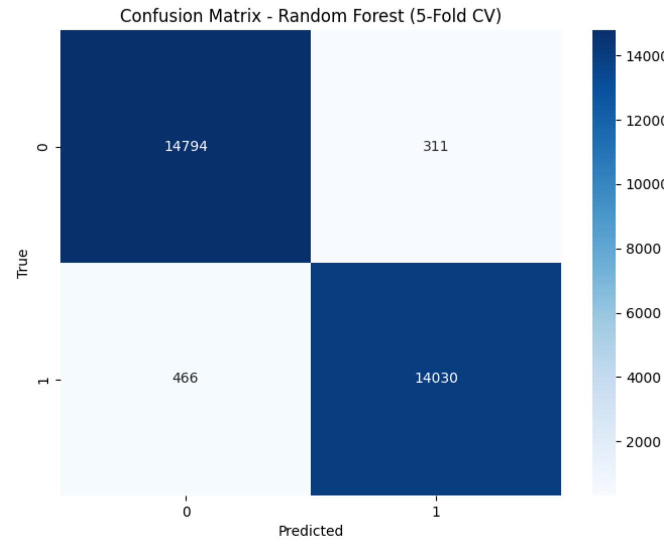
Figure 4: Top 10 most important features ranked by Random Forest model.

Table 2: Performance metrics of Random Forest model for Binary, 4-Class, and 7-Class classification.

Classification Type	Class Label	Accuracy	Precision	Recall	F1-score
Binary	Dynamic	0.98	0.97	0.98	0.98
	Static		0.98	0.97	0.98
4-Class	Walking	0.9748	0.97	0.98	0.98
	Standing		0.95	0.91	0.93
	Sitting		1.00	1.00	1.00
	Lying		1.00	1.00	1.00
7-Class	Walking	0.9524	0.94	0.98	0.96
	Shuffling		0.55	0.14	0.22
	Stairs (asc)		0.86	0.20	0.33
	Stairs (desc)		1.00	0.18	0.31
	Standing		0.93	0.92	0.93
	Sitting		1.00	1.00	1.00
	Lying		1.00	1.00	1.00
7-Class (Top 10 Features)	Walking	0.9474	0.94	0.98	0.96
	Shuffling		0.53	0.07	0.12
	Stairs (asc)		0.62	0.39	0.48
	Stairs (desc)		0.67	0.25	0.36
	Standing		0.89	0.92	0.91
	Sitting		1.00	1.00	1.00
	Lying		1.00	1.00	1.00

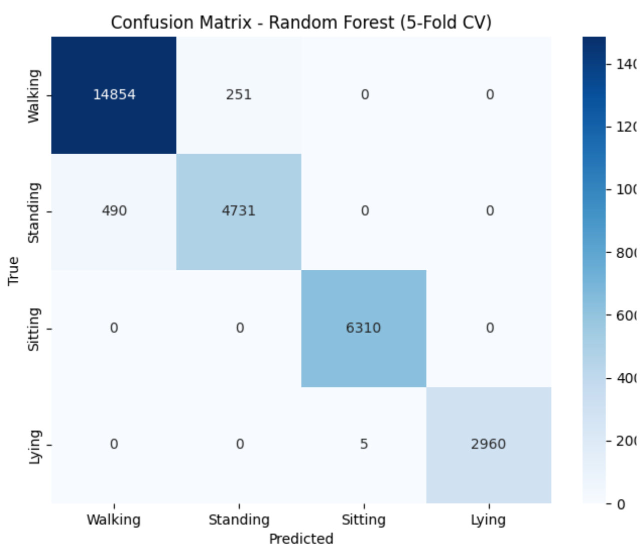


(a) Binary Classification

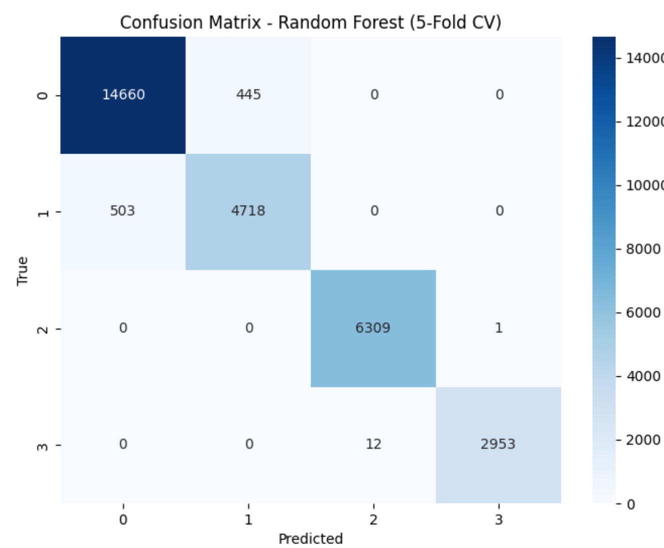


(b) 4-Class Classification

Figure 5: Confusion Matrices for Binary-Class Classification

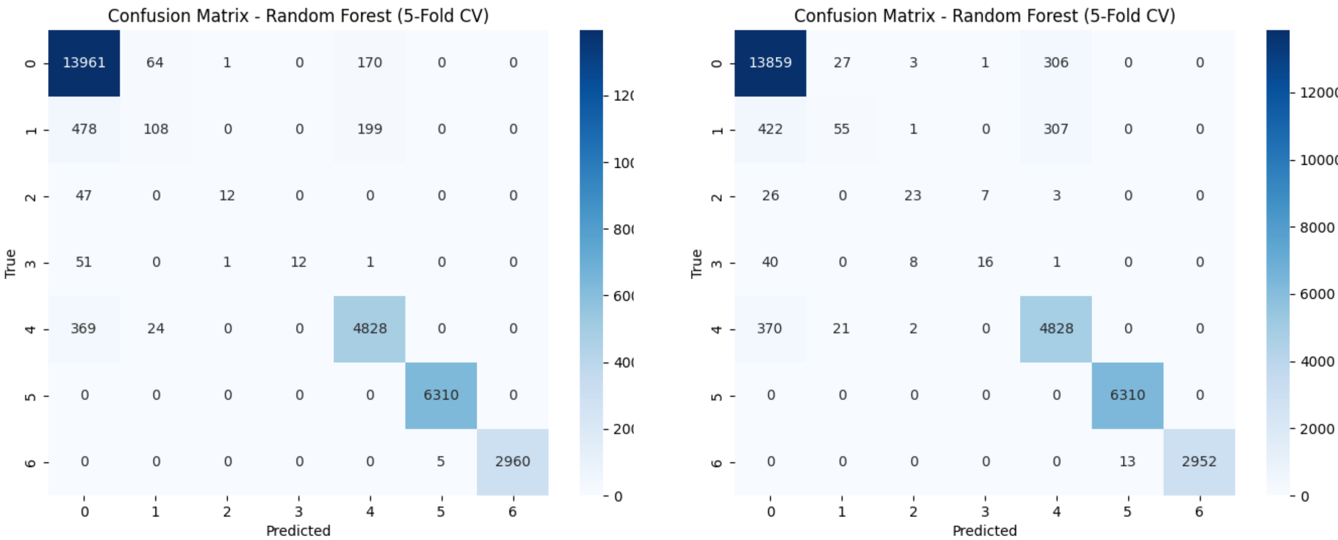


(a) Binary Classification with all features



(b) Binary Classification with top 10 features

Figure 6: Confusion Matrices for 4-Class Classification

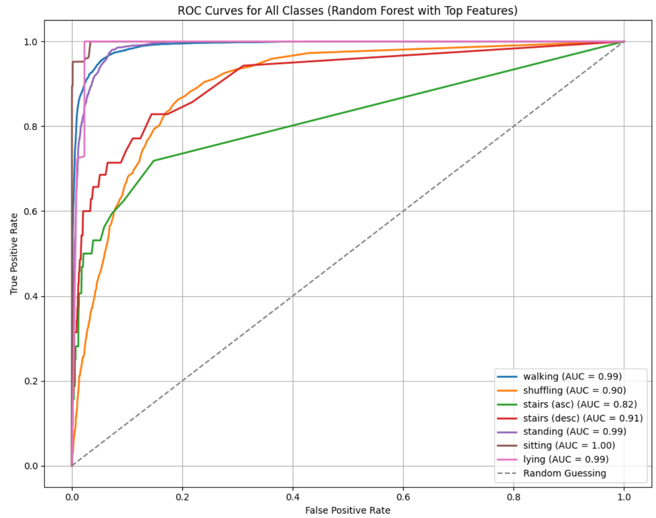


(a) 7-Class Classification with all features

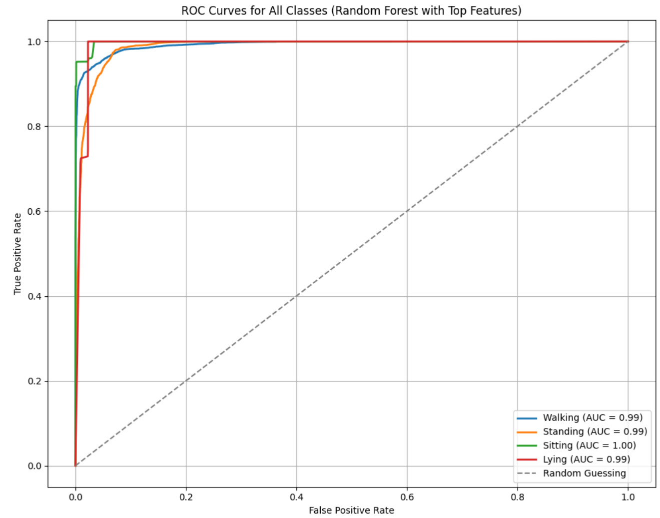
(b) 7-Class Classification with top 10 features

Figure 7: Confusion Matrices for 7-Class Classification

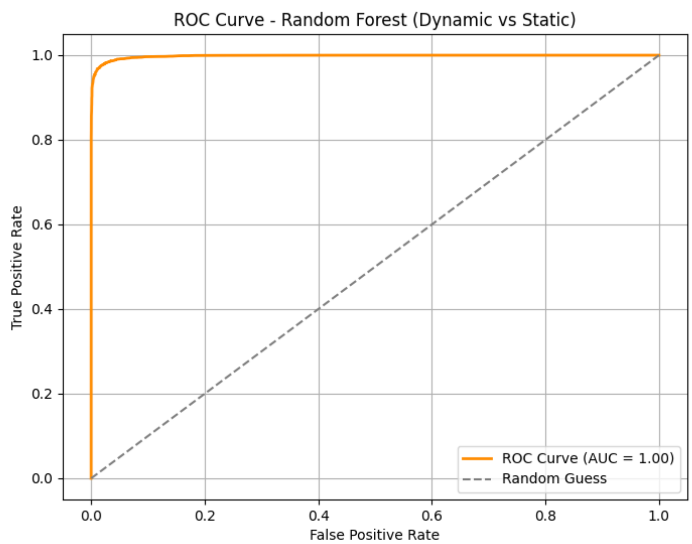
Finally, Figure 8 shows ROC curves for the Random Forest model across all three classification tasks, illustrating strong class separation and excellent AUC performance, especially in binary and 4-class settings.



(a) 7-Class Classification



(b) 4-Class Classification



(c) Binary Classification

Figure 8: ROC Curves for Binary, 4-Class, and 7-Class Classification using Random Forest.

3.3. External Validation on HARTH Dataset

To assess the generalization ability of the proposed model, we evaluated it on a subset of the **HARTH dataset** — a publicly available, professionally annotated dataset recorded in a free-living environment. Unlike HAR70+, which includes only older adults in a semi-structured setting, HARTH was collected from participants performing natural daily activities with minimal intervention. This dataset serves as a robust benchmark for testing real-world applicability.

For this external validation, we selected data from **two HARTH subjects** to simulate a light-weight, cross-domain test scenario. The same preprocessing pipeline was applied: segmentation using 1-second windows, statistical and frequency-domain feature extraction, and label mapping to the original 7-class format.

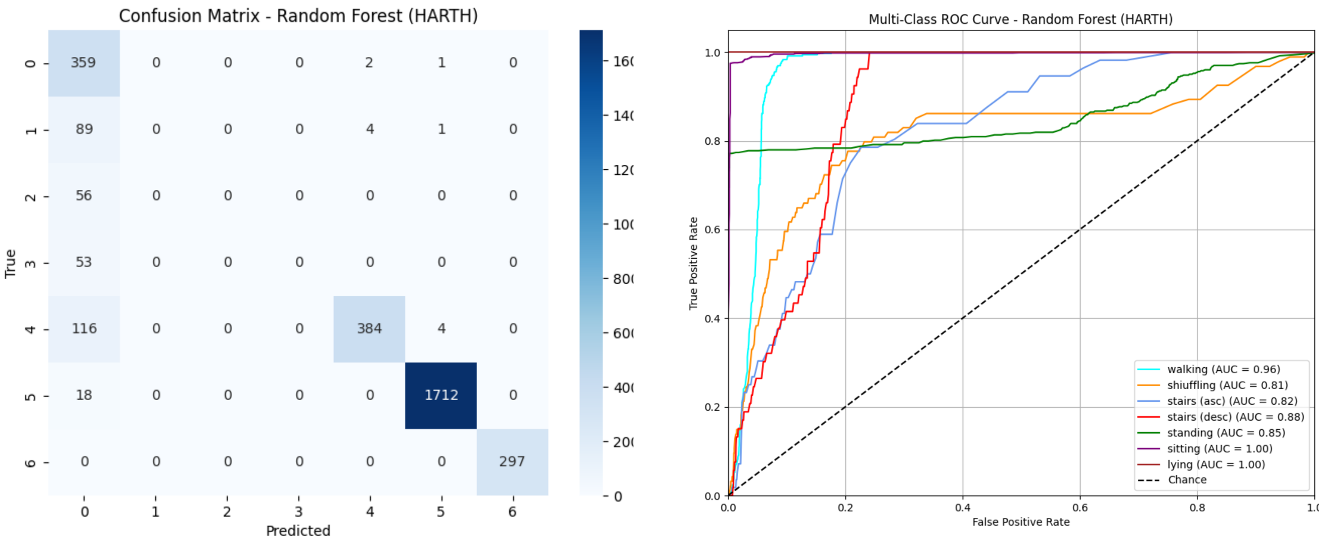
After filtering and segmentation, we extracted a total of **3,096 windows**, each represented by a 36-dimensional feature vector. These were passed to the trained Random Forest model (no retraining), simulating direct deployment on unseen data.

The model achieved an accuracy of **88.89%**. However, as shown in Table 3, performance varied across activity classes. It performed exceptionally well on static classes such as sitting and lying, but failed to recognize minority dynamic classes like stairs and shuffling, likely due to their low representation in the original training set.

Table 3: Classification report for 7-class prediction on HARTH subset (Random Forest)

Class	**Precision**	**Recall**	**F1-Score**
Walking (1)	0.52	0.99	0.68
Shuffling (3)	0.00	0.00	0.00
Stairs Ascending (4)	0.00	0.00	0.00
Stairs Descending (5)	0.00	0.00	0.00
Standing (6)	0.98	0.76	0.86
Sitting (7)	1.00	0.99	0.99
Lying (8)	1.00	1.00	1.00
Accuracy		0.8889	
Macro Average	0.50	0.53	0.50
Weighted Avg	0.87	0.89	0.87

Figure 9 visualizes the test performance through a confusion matrix and ROC curve. These highlight the model's strong capability to identify static postures, while revealing limitations in distinguishing dynamic activities not well-represented in the training set.



(a) Confusion Matrix

(b) ROC Curve

Figure 9: Evaluation on HARTH subset: confusion matrix and ROC curve for 7-class classification using Random Forest.

4. Discussion

The results from our study highlight the viability of classical machine learning approaches for accurate and interpretable Human Activity Recognition (HAR) in older adults. Through the use of handcrafted time- and frequency-domain features derived from dual accelerometer data, our Random Forest-based classifier achieved high accuracy across all classification setups—95.2% for 7-class, 97.5% for 4-class, and 98% for binary classification. These findings reinforce the utility of feature engineering in contexts where deep learning may be impractical due to limited data, hardware constraints, or the need for model interpretability.

A key insight from this study is the trade-off between classification granularity and accuracy. While the model excelled in binary and 4-class setups, performance in the 7-class scenario was hindered by class imbalance, particularly for infrequent activities such as stair ascending, descending, and shuffling. These activities had notably lower recall, suggesting the model struggled to generalize due to limited representation in the training set. This aligns with prior literature, where minority class recognition remains a persistent challenge in HAR, especially for older populations exhibiting atypical movement patterns.

The confusion matrices and class-wise metrics clearly show that static activities such as sitting and lying were classified with near-perfect accuracy. This may be attributed to their consistent and distinguishable sensor signatures. Conversely, dynamic and transitional activities like shuffling and stair navigation were often misclassified, possibly due to overlap in movement patterns or insufficient variation in the training data. While the use of `class_weight='balanced'` helped mitigate this issue to some extent, further improvement might require targeted oversampling, data augmentation, or advanced techniques such as synthetic data generation from motion models.

Our evaluation on the HARTH dataset provides encouraging evidence of model generalization, achieving 88.9% accuracy despite differences in participant behavior, environment, and data distribution. However, external performance still suffered on underrepresented dynamic classes—indicating that model robustness across domains depends on the diversity and coverage of the training data. This suggests a need for larger, more heterogeneous datasets or domain adaptation strategies to ensure wide applicability.

Feature selection also proved valuable; a reduced 10-feature model retained high performance, offering a path to efficient on-device deployment. The minor drop in accuracy from full-feature to top-feature models indicates that a small, well-chosen feature set can preserve performance while reducing computational load.

Despite these strengths, limitations of the current work include the small sample size (18 subjects) and reliance on data from only two sensor locations. Although we focused on interpretability and real-time feasibility, this may limit the system's sensitivity to finer movement nuances. Future work could explore multi-modal sensing (e.g., combining gyroscopes or pressure sensors), unsupervised pretraining, or personalized models that adapt to user-specific movement patterns.

To contextualize our results, Table 4 compares our approach with recent HAR70+ studies. Our model outperformed or matched others in accuracy while maintaining a simpler and more interpretable feature set. Deep learning approaches such as ResBiGRU and LSTM-GRU offered slightly higher performance but at the cost of complexity and reduced interpretability.

Table 4: Comparison of Proposed Model with Related HAR70+ Studies

Study	**Dataset**	**Features Used**	**Activities Recognized**	**Classifier**	**Accuracy**
Astrid et al. [41]	HAR70+	Top 10 from 161 (Stat. + Freq. + Movement + Cross-sensor)	Walking, Standing, Sitting, Lying	XGBoost	93%
Sidra et al. [42]	HAR70+	Recursive Feature Elimination + PCA	All 7 Activities	Random Forest	93.95%
Heba et al. [43]	HAR70+	Mean, RMS, Variance, Energy-to-Shannon Ratio	All 7 Activities	Decision Tree	84.64%
Sakorn et al. [44]	HAR70+	Deep Learning-Based Method	All 7 Activities	ResBiGRU	97.39%
Albogamy et al. [45]	HAR70+	Deep Learning-Based Method	All 7 Activities	Hybrid LSTM-GRU	95.2%
Proposed Study	HAR70+	Top 10 from 36 (Stat. + FFT)	All 7 Activities	Random Forest	**95.2%**

5. Conclusion

In this study, we developed an efficient and interpretable machine learning framework for recognizing human activities in older adults using dual accelerometer data. By extracting time- and frequency-domain features from segmented sensor windows and employing a Random Forest classifier with extensive hyperparameter tuning, the system achieved high accuracy across multiple classification tasks, including fine-grained and simplified setups. The model demonstrated strong generalization on external data, validating its applicability in real-world scenarios. Importantly, the use of a reduced feature set preserved performance while enabling lightweight deployment. These findings underscore the potential of classical machine learning with handcrafted features as a practical solution for activity monitoring in aging populations, especially where computational efficiency, interpretability, and robustness are critical.

References

[1] K. R. Evenson, E. Scherer, K. M. Peter, C. C. Cuthbertson, S. Eckman, Historical development of accelerometry measures and methods for physical activity and sedentary behavior research worldwide: A scoping review of observational studies of adults, PLoS One 17 (11) (2022) e0276890.

[2] M. L. Stevens, N. Gupta, E. I. Eroglu, P. J. Crowley, B. Eroglu, A. Bauman, M. Granat, L. Straker, P. Palm, S. Stenholm, et al., Thigh-worn accelerometry for measuring movement and posture across the 24-hour cycle: a scoping review and expert statement, BMJ Open Sport & Exercise Medicine 6 (1) (2020).

[3] K. Wijndaele, K. Westgate, S. K. Stephens, S. N. Blair, F. C. Bull, S. F. Chastin, D. W. Dunstan, U. Ekelund, D. W. Esliger, P. S. Freedson, et al., Utilization and harmonization of adult accelerometry data: review and expert consensus, Medicine and science in sports and exercise 47 (10) (2015) 2129.

[4] F. Alam, A. Almaghthawi, I. Katib, A. Albeshri, R. Mehmood, Iresponse: An ai and iot-enabled framework for autonomous covid-19 pandemic management, Sustainability 13 (7) (2021) 3797.

[5] A. Ganesan, A. Paul, H. Seo, Elderly people activity recognition in smart grid monitoring environment, Mathematical Problems in Engineering 2022 (1) (2022) 9540033.

[6] S. M. Laxmi Pai, Child activity recognition using accelerometer and rfid reader.

[7] W. A. Kusuma, A. E. Minarno, N. D. N. Safitri, Human activity recognition utilizing svm algorithm with gridsearch, in: AIP Conference Proceedings, Vol. 2453, AIP Publishing, 2022.

[8] H. Qian, Y. Mao, W. Xiang, Z. Wang, Recognition of human activities using svm multi-class classifier, Pattern Recognition Letters 31 (2) (2010) 100–111.

[9] K. M. Chathuramali, R. Rodrigo, Faster human activity recognition with svm, in: International conference on advances in ICT for emerging regions (ICTer2012), IEEE, 2012, pp. 197–203.

[10] D. Anguita, A. Ghio, L. Oneto, X. Parra, J. L. Reyes-Ortiz, Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine, in: Ambient Assisted Living and Home Care: 4th International Workshop, IWAAL 2012, Vitoria-Gasteiz, Spain, December 3-5, 2012. Proceedings 4, Springer, 2012, pp. 216–223.

[11] Z. Feng, Y. Shi, D. Zhou, L. Mo, Research on human activity recognition based on random forest classifier, in: 2023 IEEE International Conference on Control, Electronics and Computer Technology (ICCECT), IEEE, 2023, pp. 1507–1513.

[12] L. Xu, W. Yang, Y. Cao, Q. Li, Human activity recognition based on random forests, in: 2017 13th international conference on natural computation, fuzzy systems and knowledge discovery (ICNC-FSKD), IEEE, 2017, pp. 548–553.

[13] Y. Zhang, X. Zhao, Z. Li, Facilitated and enhanced human activity recognition via semi-supervised lightgbm, in: 2020 IEEE Globecom Workshops (GC Wkshps, IEEE, 2020, pp. 1–6.

[14] Z. Shao, J. Guo, Y. Zhang, R. Zhu, L. Wang, Lightbgm for human activity recognition using wearable sensors, in: 2021 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS), IEEE, 2021, pp. 668–671.

[15] G. Csizmadia, K. Liszkai-Peres, B. Ferdinandy, Á. Miklósi, V. Konok, Human activity recognition of children with wearable devices using lightgbm machine learning, Scientific Reports 12 (1) (2022) 5472.

[16] K. Maswadi, N. A. Ghani, S. Hamid, M. B. Rasheed, Human activity classification using decision tree and naïve bayes classifiers, Multimedia Tools and Applications 80 (2021) 21709–21726.

[17] J. Shen, H. Fang, Human activity recognition using gaussian naïve bayes algorithm in smart home, in: Journal of Physics: Conference Series, Vol. 1631, IOP Publishing, 2020, p. 012059.

[18] Z. Liu, S. Li, J. Hao, J. Hu, M. Pan, An efficient and fast model reduced kernel knn for human activity recognition, Journal of Advanced Transportation 2021 (1) (2021) 2026895.

[19] S. Mohsen, A. Elkaseer, S. G. Scholz, Human activity recognition using k-nearest neighbor machine learning algorithm, in: Proceedings of the International Conference on Sustainable Design and Manufacturing, Springer, 2021, pp. 304–313.

[20] C. A. Ronao, S.-B. Cho, Human activity recognition with smartphone sensors using deep learning neural networks, Expert systems with applications 59 (2016) 235–244.

[21] S. Ha, S. Choi, Convolutional neural networks for human activity recognition using multiple accelerometer and gyroscope sensors, in: 2016 international joint conference on neural networks (IJCNN), IEEE, 2016, pp. 381–388.

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

[23] M. Zeng, L. T. Nguyen, B. Yu, O. J. Mengshoel, J. Zhu, P. Wu, J. Zhang, Convolutional neural networks for human activity recognition using mobile sensors, in: 6th international conference on mobile computing, applications and services, IEEE, 2014, pp. 197–205.

[24] F. J. O. Morales, D. Roggen, Deep convolutional feature transfer across mobile activity recognition domains, sensor modalities and locations, in: Proceedings of the 2016 ACM International Symposium on Wearable Computers, 2016, pp. 92–99.

[25] Y. Guan, T. Plötz, Ensembles of deep lstm learners for activity recognition using wearables, Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies 1 (2) (2017) 1–28.

[26] M. Awais, L. Chiari, E. A. F. Ihlen, J. L. Helbostad, L. Palmerini, Physical activity classification for elderly people in free-living conditions, IEEE journal of biomedical and health informatics 23 (1) (2018) 197–207.

[27] J. Cherian, V. Rajanna, D. Goldberg, T. Hammond, Did you remember to brush? a noninvasive wearable approach to recognizing brushing teeth for elderly care, in: Proceedings of the 11th EAI International Conference on Pervasive Computing Technologies for Healthcare, 2017, pp. 48–57.

[28] N. Noury, P. Barralon, N. Vuillerme, A. Fleury, Fusion of multiple sensors sources in a smart home to detect scenarios of activities in ambient assisted living, International Journal of E-Health and Medical Communications (IJEHMC) 3 (3) (2012) 29–44.

[29] A. Mihoub, A deep learning-based framework for human activity recognition in smart homes, Mobile Information Systems 2021 (1) (2021) 6961343.

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

[31] G. Dogan, S. S. Ertas, I. Cay, Human activity recognition using convolutional neural networks, in: 2021 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), IEEE, 2021, pp. 1–5.

[32] I. U. Khan, S. Afzal, J. W. Lee, Human activity recognition via hybrid deep learning based model, Sensors 22 (1) (2022) 323.

[33] K. M. L. Durga, P. Jyotsna, G. K. Kumar, A deep learning based human activity recognition model using long short term memory networks, in: 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), IEEE, 2022, pp. 1371–1376.

[34] E. S. Rahayu, E. M. Yuniarno, I. K. E. Purnama, M. H. Purnomo, Human activity classification using deep learning based on 3d motion feature, Machine Learning with Applications 12 (2023) 100461.

[35] F. Alam, R. Mehmood, I. Katib, A. Albeshri, Analysis of eight data mining algorithms for smarter internet of things (iot), Procedia Computer Science 98 (2016) 437–442.

[36] T. Yu, H. Zhu, Hyper-parameter optimization: A review of algorithms and applications, arXiv preprint arXiv:2003.05689 (2020).

[37] C. Cooney, A. Korik, R. Folli, D. Coyle, Evaluation of hyperparameter optimization in machine and deep learning methods for decoding imagined speech eeg, Sensors 20 (16) (2020) 4629.

[38] Y. J. H. Z. W. L. J. L. J. Z. C. C. B. Li, Yuwei; Shen, Hyper-tune: Towards efficient hyper-parameter tuning at scale, arXiv preprint arXiv:2201.06834 (2022). `arXiv:2201.06834`.

[39] B. Bischl, M. Binder, M. Lang, T. Pielok, J. Richter, S. Coors, J. Thomas, T. Ullmann, M. Becker, A.-L. Boulesteix, et al., Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 13 (2) (2023) e1484.

[40] M. A. K. Raiaan, S. Sakib, N. M. Fahad, A. Al Mamun, M. A. Rahman, S. Shatabda, M. S. H. Mukta, A systematic review of hyperparameter optimization techniques in convolutional neural networks, Decision analytics journal (2024) 100470.

[41] A. Ustad, A. Logacjov, S. Ø. Trollebø, P. Thingstad, B. Vereijken, K. Bach, N. S. Maroni, Validation of an activity type recognition model classifying daily physical behavior in older adults: the har70+ model, Sensors 23 (5) (2023) 2368.

[42] S. Abbas, G. A. Sampedro, S. Alsubai, S. Ojo, A. S. Almadhor, A. Al Hejaili, L. Strazovska, Advancing healthcare and elderly activity recognition: Active machine and deep learning for fine-grained heterogeneity activity recognition, IEEE Access 12 (2024) 44949–44959.

[43] H. Nematallah, S. Rajan, Quantitative analysis of mother wavelet function selection for wearable sensors-based human activity recognition, Sensors 24 (7) (2024) 2119.

[44] S. Mekruksavanich, W. Phaphan, A. Jitpattanakul, Harnessing deep learning for activity recognition in seniors' daily routines with wearable sensors, in: 2024 47th International Conference on Telecommunications and Signal Processing (TSP), IEEE, 2024, pp. 164–167.

[45] F. R. Albogamy, Federated learning for iomt-enhanced human activity recognition with hybrid lstm-gru networks, Sensors 25 (3) (2025) 907.