PENINSULA PUBLISHING LLC

EDRAAK
Vol. (2026), 2026, pp. 22–40
ISSN: 3078-8412

Research Article

Human Activity Recognition Using Smartphone Sensor Data: A Lightweight Benchmark Study

Abdulazeez Alsajri[1, *], Amani Steiti[2],

[1] *Computer Science Department, University Arts, Sciences and Technology, Beirut, Lebanon*

[2] *Department of Networks and Computer Systems, Faculty of Informatics Engineering, Al Wataniya Private University, Hama, Syria*

ARTICLEINFO

Article History

Received 5 Dec 2025

Revised: 29 Jan 2026
Accepted 1 Mar 2026
Published 15 Mar 2026

Keywords

Human Activity Recognition,

Smartphone Sensors,

Lightweight Benchmark,

Machine Learning,

Activity Classification,

Feature Engineering,

Principal Component Analysis,

Mobile Computing.

ABSTRACT

Smartphone inertial sensors enable Human Activity Recognition (HAR) which serves as a fundamental task for mobile computing and digital health and context-aware systems because these devices offer both functional sensing abilities and extensive access to real-world environments. The study creates a journal-style lightweight benchmark for smartphone sensor-based HAR through a dataset which contains 4,999 samples and 561 numerical features and 15 subjects and six activity classes after data cleaning and restructuring. The proposed pipeline performs data repair before preprocessing and descriptive statistics and feature grouping and correlation analysis and principal component analysis (PCA) and ANOVA-based feature ranking and model benchmarking. The study tested four lightweight machine learning baselines which included Logistic Regression and Random Forest and k-Nearest Neighbors (KNN) and Gaussian Naive Bayes through an 80/20 train-test split that maintained class distribution. The Logistic Regression model delivered the highest performance with an accuracy of 0.973 and a macro F1-score of 0.9748 yet Random Forest and KNN models also achieved results above 0.95. The analysis of features demonstrated that gravity-based descriptors together with angle-based descriptors produced the most effective recognition results yet the error analysis showed that the most frequent classification errors occurred between sitting and standing positions. The research demonstrates that feature representations which have been designed with purpose maintain their competitive position when used in deployment-focused HAR systems.

1. INTRODUCTION

Human Activity Recognition (HAR) systems use sensor data to identify human movement patterns which have evolved into a well-established field within mobile and ubiquitous computing. Smartphone-based HAR is especially important because smartphones are already equipped with inertial sensors, are nearly ubiquitous, and can support a wide range of applications including digital health, assisted living, rehabilitation, sports analytics, and context-aware services [1]– [4]. The smartphone sensors which include accelerometers and gyroscopes serve as the most useful data acquisition tools for operational HAR systems according to multiple reviews published between 2021 and 2024 [1]– [3].

Research in this field has evolved through different established stages of development. Researchers in the beginning stages demonstrated that mobile devices used by consumers could identify their daily activities through basic machine learning techniques [6]–[9]. The availability of public benchmark datasets including the UCI smartphone HAR dataset enabled researchers to perform experiments which could be duplicated by others while creating an environment for unbiased assessment of different models [6], [7]. Deep learning methods which use CNN and LSTM and hybrid CNN-LSTM models focus on automatic feature extraction from unprocessed or slightly processed sensor data [4], [5], [9]– [11]. The field now focuses on developing small models which operate at maximum efficiency for edge and mobile devices according to studies from 1, 13, 16 through 19.

The transformation creates a vital shift in the situation. The best HAR model for actual use goes beyond achieving top accuracy in controlled settings because it achieves the best balance between discrimination and latency and memory cost and interpretability and deployment feasibility. The trade-off between performance and power consumption and computational requirements proves crucial when operating in smartphone and edge environments because these systems

*Corresponding author email: aka104@live.aul.edu.lb
DOI: https://doi.org/10.70470/EDRAAK/2026/003

need to maintain their real-time functionality and battery life and efficient use of computational resources [13], [16]– [19]. Researchers continue to focus on basic accuracy because they view it as their primary source of financial support. Real systems do not operate according to romantic principles.

The current research establishes a lightweight benchmark document which uses a feature-rich smartphone HAR dataset that the user provided and then processed through cleaning and organization and analysis and comparison in a single system. The paper has four main objectives: (i) to prepare a publication-ready description of the dataset and preprocessing pipeline, (ii) to conduct exploratory and discriminative feature analysis, (iii) to benchmark several lightweight classical models on the processed data, and (iv) to interpret the resulting performance in light of current literature on smartphone HAR and deployment-efficient modeling.

The main research questions for this study include RQ1 which investigates the performance of different lightweight classical models regarding their ability to achieve both high accuracy and efficient processing on the smartphone HAR dataset. RQ2—Which feature families appear most informative for the recognition of the six selected activity classes? RQ3—Does a feature-based lightweight benchmark still provide competitive and practically meaningful results in an era dominated by deep learning? The paper answers these questions using a complete article structure and an evidence-backed discussion.

2. RELATED WORK

Researchers have been investigating smartphone-based Human Activity Recognition (HAR) systems since the early 2000s when foundational research proved that mobile inertial data could identify common human activities. Kwapisz et al. demonstrated that cell phone accelerometers could effectively support activity classification using conventional machine learning methods [8]. The public benchmark dataset released by Anguita et al. became a leading resource which smartphone HAR researchers use to evaluate their methods through standardized comparison techniques [6], [7]. The WISDM dataset together with smartphone-smartwatch data established a new experimental domain which researchers used to investigate device differences and expanded activity categories and combined multiple sensing technologies [20].

Research shows that smartphone-based Human Activity Recognition (HAR) has expanded beyond its original restricted application domain. Morales and Akopian examined how smartphone-based HAR systems operate through their analysis of sensing technologies and placement methods and preprocessing techniques and feature extraction methods and usability challenges [3]. The research by Wang et al. and Nweke et al. explored how deep learning techniques for sensor-based activity recognition developed through time while showing that convolutional and recurrent and hybrid network structures decreased the need for manual feature engineering but required more complex network designs [4], [5]. The survey by Dent amaro et al. investigated smartphone sensor integration to show present methods for data preprocessing and feature engineering and selection and classification techniques [1][2]. The research by Ferrari et al. identified major developments which emerged through smartphone HAR system operations and their used datasets and evaluation methodologies [13].

Deep learning brought a complete transformation to the entire field. Ronao and Cho demonstrated that deep convolutional networks enabled direct learning of strong activity representations from smartphone sensor data [9]. Ordóñez and Roggen developed their research field through a method which merges convolutional operations with LSTM networks for time-based analysis [10]. The research community found CNN and LSTM and CNN-LSTM and TCN and attention-based and transformer-inspired methods to perform well in their studies when they processed raw window data as input [4], [5], [11], [16]–[19], [21], [22].

At the same time, lightweight modeling has become a serious design target rather than a side note. The authors Agarwal and Alam developed a lightweight HAR system which operates on edge devices [19]. The authors Sekaran et al. developed Light-MHTCN which operates as a lightweight multiheaded temporal convolutional network to perform smartphone-based HAR [16]. The research team at Gong et al. developed MobileHARC and Wang et al. created MobileHAR through inverted residual inception blocks to achieve decreased computational requirements while preserving strong recognition results [18], [23]. The 2024 lightweight deep learning study by AlMuhaideb et al. follows the same research direction [17].

The progress in technology has not eliminated the necessity for benchmark systems which base their evaluations on clear features. The majority of actual implementations continue to use systems which combine understandable processing methods with manually developed features and predictive models that run efficiently. The understanding of discriminative structures together with redundancy and signal family contributions to recognition performance becomes possible through the use of feature-rich datasets [1], [3], [13], [15]. The paper serves as a detailed benchmark study which demonstrates the potential of disciplined data preparation methods when combined with traditional lightweight learning algorithms instead of competing with contemporary compact deep models[14].

3. METHODOLOGY

3.1 Dataset Description and Cleaning

The uploaded dataset contained a structurally corrupted CSV header that initially prevented direct analysis. The data itself, however, was intact. The dataset became usable after I fixed its header and converted feature columns into numeric format which resulted in 4,999 samples and 561 numerical features and 15 subjects and 6 activity classes. The activity labels were LAYING, SITTING, STANDING, WALKING, WALKING_DOWNSTAIRS, and WALKING_UPSTAIRS. The dataset became complete because it contained no missing values and all duplicate records were removed during the cleaning process [12].

The feature set uses the UCI smartphone HAR benchmark approach which applies a sliding-window system to create time-domain and frequency-domain descriptors from accelerometer and gyroscope data [6], [7]. The current dataset includes two final metadata columns which contain subject identification and activity labeling and all other columns contain features which we generated from different signal types.

TABLE I. DATASET OVERVIEW

Attribute	Value
Samples	4999
Features	561
Activity classes	6
Subjects	15
Missing values	0
Duplicate rows	0

The table provides fundamental details about the cleaned dataset through its display of sample counts and feature space dimensions and activity class totals and subject numbers and confirmation of record completeness without any duplicates or missing values. The benchmark documentation provides a methodological function because it demonstrates the benchmark derives from a well-organized dataset which contains no missing data or noise.

TABLE II. ACTIVITY DISTRIBUTION.

Activity	Count	Percentage
LAYING	965	19.3000
STANDING	943	18.8600
SITTING	879	17.5800
WALKING	850	17.0000
WALKING_UPSTAIRS	711	14.2200
WALKING_DOWNSTAIRS	651	13.0200

The table displays how often each activity class appears together with its respective percentage distribution. The classification balance between classes creates essential conditions to properly evaluate both accuracy and F1-score because a balanced class distribution between them prevents majority-class data from creating false impressions of model performance.

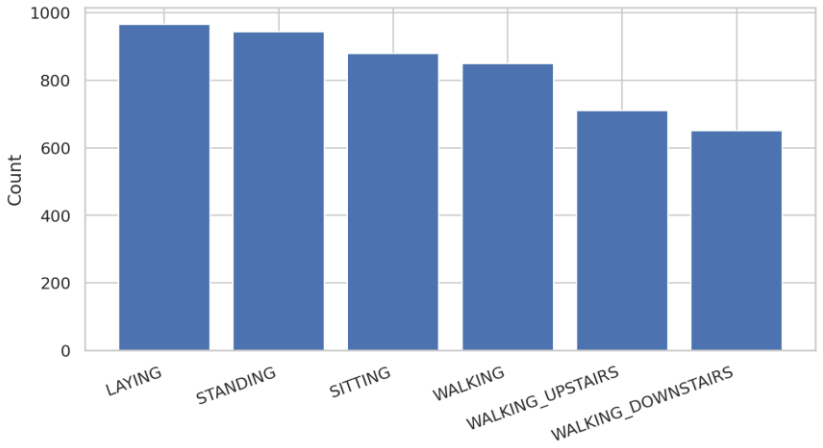

Fig. 1. Distribution of activity classes in the cleaned dataset.

This figure visualizes the class counts reported in Table II. The bar heights make class balance easier to assess at a glance and help the reader see that the benchmark is not dominated by a single activity class.

TABLE III. SUBJECT DISTRIBUTION.

Subject	Count
1	347
3	341
5	302
6	325
7	308
8	281
11	316
14	323
15	328
16	366
17	368
19	360
21	408
22	321
23	305

The table displays the allocation of samples which were collected from various participants in the study. The distribution of subjects becomes essential because their concentration will create biased activity models through their excessive influence in the dataset. The subject numbers in this location are well distributed to create a reliable benchmark for analysis.

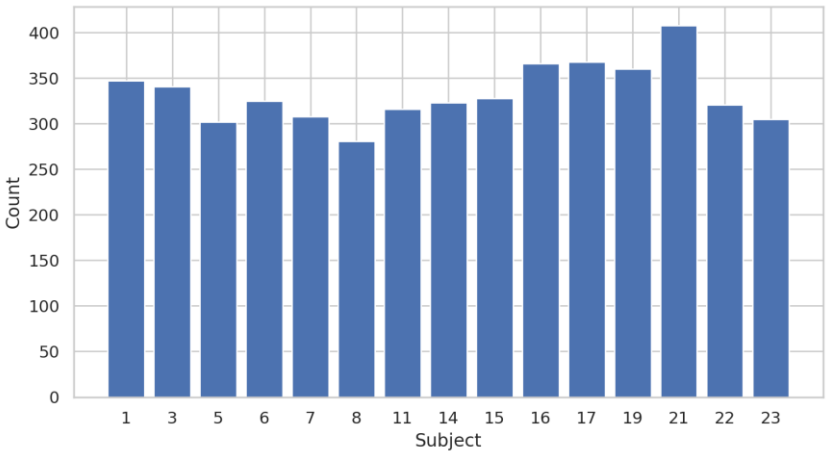

Fig. 2. Number of samples contributed by each subject.

The diagram shows how each subject participated in the study which extends the data presented in Table III. The diagram shows how different subjects contributed to the work which enables users to identify which participants could distort the benchmark through their limited representation.

3.2 Feature Organization and Preprocessing

The study of feature organization focused on developing understandable signal groups which included multiple predictors. The dataset included multiple data types which consisted of time-domain body acceleration and gravity acceleration and body acceleration jerk and body gyroscope and gyroscope jerk and magnitude-based measures and frequency-domain measures and angle-based features. The process of feature grouping fulfills two essential functions because it enhances data comprehension while allowing us to organize our upcoming feature analysis discussions effectively.

Data processing required three main steps which included label normalization and subject-field standardization and feature matrix z-score scaling before applying dimensionality reduction and modeling techniques. The process of standardization serves as a crucial step because Logistic Regression and KNN classifiers which were benchmarked require data to be scaled for proper operation. The scaled feature matrix was also used for PCA-based variance analysis and low-dimensional projection.

TABLE IV. FEATURE-GROUP SUMMARY.

Feature group	Count
Time-domain body acceleration	40
Time-domain gravity acceleration	40
Time-domain body acceleration jerk	40
Time-domain body gyroscope	40
Time-domain gyroscope jerk	40
Magnitude features	117
Frequency-domain features	289
Angle features	7

The table presents 561 predictors which have been categorized into four main signal families that include time-domain descriptors and frequency-domain descriptors and magnitude descriptors and angle-based descriptors. The benchmark processes structured sensor-based features which contain specific physical meanings instead of working with unidentified columns according to the grouping system which aids reader understanding.

$$z = (x - \mu) / \sigma \qquad (1)$$

Equation 1. Z-Score Standardization

Equation (1) standardizes each feature by centering it at the mean and scaling it by the standard deviation. The learning process avoids being dominated by variables which have large numeric ranges because this method implements a prevention system that works particularly well with distance-based and linear classification methods.

$$\mu = (1/N) \Sigma x_i \qquad (2)$$

Equation 2. Arithmetic Mean

Equation (2) defines the arithmetic mean, which summarizes the central tendency of a feature. The concept finds application in descriptive statistics and also serves as a fundamental component for calculating standardization and variance values.

$$\sigma = \sqrt{[(1/N) \Sigma (x_i - \mu)^2]} \qquad (3)$$

Equation 3. Standard Deviation

Equation (3) defines the standard deviation as a measure of dispersion around the mean. The research requires this method because it enables basic data evaluation and primary data transformation before processing the feature matrix during the initial stages of data analysis.

$$Var(x) = (1/N) \Sigma (x_i - \mu)^2 \qquad (4)$$

Equation 4. Variance

Equation (4) shows how to calculate variance which determines the typical squared distance between data points and their mean value. Variance serves as an analytical tool because features with high variance tend to represent major signal changes although they do not necessarily provide the best class separation.

$$SMA = (1/N) \Sigma (|x_i| + |y_i| + |z_i|) \qquad (5)$$

Equation 5. Signal Magnitude Area (SMA)

Equation (5) defines Signal Magnitude Area as a standard inertial-sensing descriptor which calculates the total magnitude of three-dimensional motion data. The method serves as a standard practice for HAR because it measures total movement strength which does not depend on specific movement directions.

3.3 Exploratory Analysis, Dimensionality Reduction, and Feature Ranking

The exploratory analysis process helped us identify how features spread across data and their redundant relationships and the organization of different classes. Table V contains the descriptive statistics which show the values of the representative variables. The top twenty features which showed the most variation were used to create a correlation map that displayed their redundant relationships and their linked features. The standardized matrix underwent PCA analysis to create a visual representation of class patterns and to measure the total amount of variance which PCA components explained.

ANOVA F-scores helped us determine which features had the strongest ability to differentiate between the six different activity categories. The top-variance features emerged from the calculation process which determined the highest dispersion predictors. The main features from the ANOVA analysis received their mean values for each activity to assist with identifying patterns which are specific to each class.

TABLE V. DESCRIPTIVE STATISTICS FOR REPRESENTATIVE FEATURES.

Feature	Mean	Std	Min	Max
tBodyAcc-mean()-X	0.2737	0.0703	-1.0000	0.6315
tBodyAcc-std()-X	-0.5953	0.4623	-1.0000	1.0000
tGravityAcc-mean()-X	0.6663	0.4988	-0.9367	0.9915
tBodyGyro-std()-Y	-0.6682	0.3806	-1.0000	1.0000
fBodyAccJerk-entropy()-X	-0.2696	0.7613	-1.0000	1.0000

The table presents combined data about typical values and value spread for main predictive variables in the dataset. The method shows basic distribution characteristics which include data range and variability throughout all features which motivates data standardization and identifies features with potential discrimination patterns.

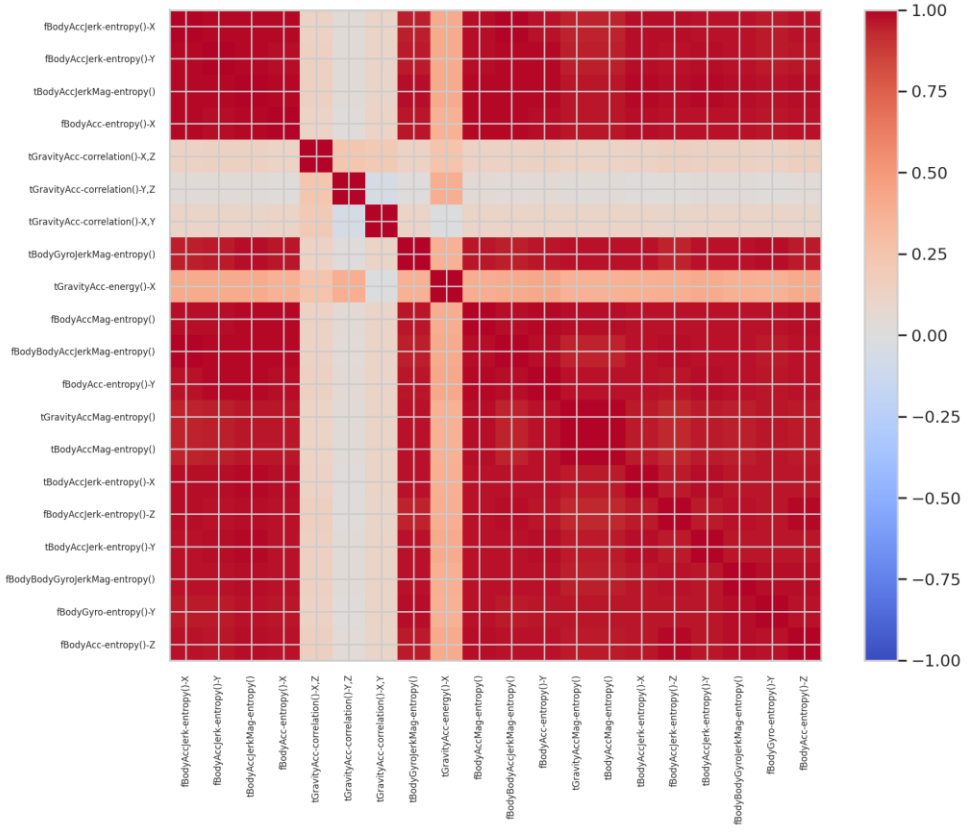

Fig. 3. Correlation heatmap for the 20 most variable features.

The diagram shows how the predictors with the highest variance values connect to each other. The method enables users to analyze multicollinearity between features in their data because it helps them determine which methods for feature selection and dimensionality reduction will work effectively.

$$Z = XW \qquad (6)$$

Equation 6. Principal Component Analysis Transformation

Equation (6) shows the PCA transformation which implements dimensionality reduction of the original feature matrix by projecting it onto an orthogonal basis of lower dimensions. The manuscript employs PCA for visual representation and variance analysis instead of using it as the ultimate classification framework.

TABLE VI. PCA EXPLAINED VARIANCE FOR THE FIRST TEN PRINCIPAL COMPONENTS.

PC	Explained Variance Ratio	Cumulative Variance
PC1	0.5081	0.5081
PC2	0.0617	0.5698
PC3	0.0292	0.5990
PC4	0.0262	0.6251
PC5	0.0194	0.6445
PC6	0.0187	0.6633
PC7	0.0143	0.6776
PC8	0.0128	0.6904
PC9	0.0105	0.7009
PC10	0.0103	0.7112

The table displays the amount of variance which the main principal components manage to capture. The analysis aims to determine if the initial 561-dimensional data contains excessive redundancy and which condensed dimensional space retains the majority of its structural information.

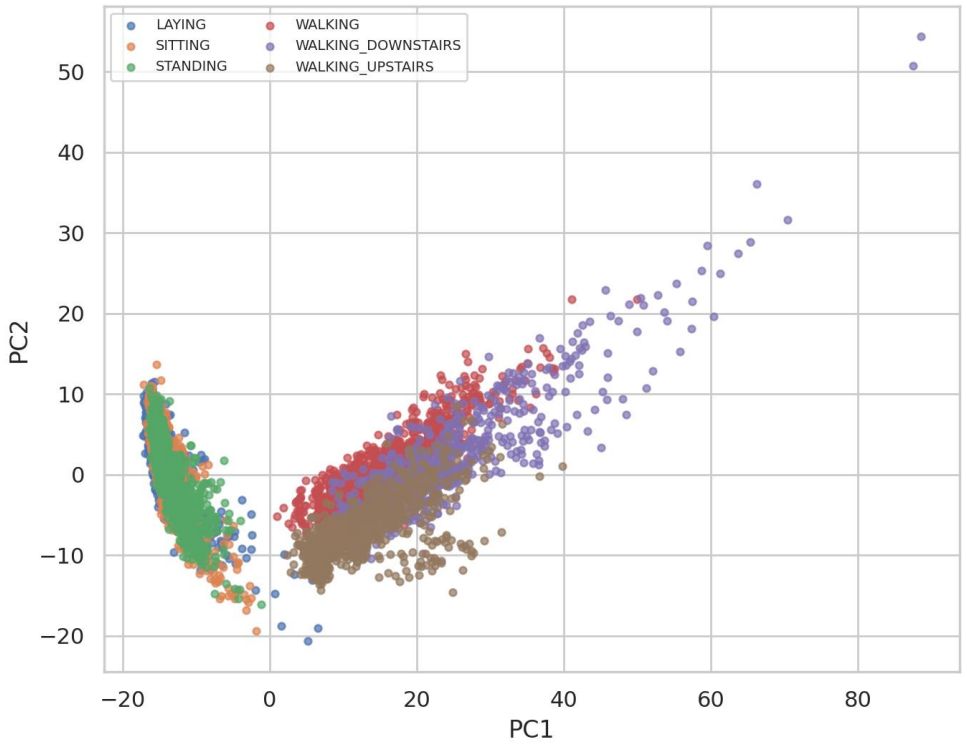

Fig. 4. Two-dimensional PCA projection of the six activity classes.

The diagram shows how the complete set of features gets mapped into two main components which reveal the visual structure of different classes. The two-dimensional projection does not keep every detail but it shows cluster distribution and activity intersections in a way that is simple to understand.

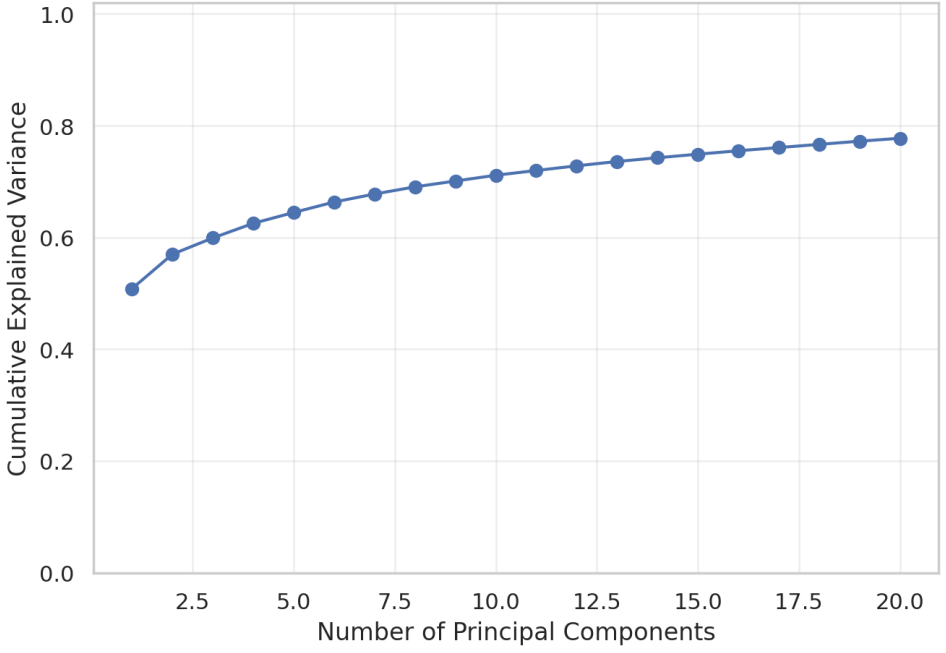

Fig. 5. Cumulative explained variance across the first 20 principal components.

The graph demonstrates the pace at which preserved variance increases when you include additional principal components. The method evaluates the compressibility of the high-dimensional feature space and identifies which orthogonal directions contain most of the information in the data.

TABLE VII. TOP FEATURES RANKED BY ANOVA F-SCORE.

Feature	ANOVA F-score	p-value
fBodyAccJerk-entropy()-X	26878.03	0
fBodyAccJerk-entropy()-Y	19624.35	0
tGravityAcc-mean()-X	18647.51	0
fBodyBodyAccJerkMag-entropy()	18069.25	0
tGravityAcc-min()-X	17980.33	0
tGravityAcc-energy()-X	17111.53	0
tGravityAcc-max()-X	16494.04	0
tBodyAcc-max()-X	15991.95	0
fBodyAcc-entropy()-X	15731.72	0
tBodyAccJerkMag-entropy()	15691.00	0

The table displays the features which create the biggest distinction between different classes when compared to the variation present within each class. The table displays the statistical performance of engineered predictors to identify which ones perform best at recognizing human activities between six different types of activities.

TABLE VIII. ACTIVITY-WISE MEANS FOR THE TOP ANOVA-RANKED FEATURES.

Activity	fBodyAccJerk-entropy()-X	fBodyAccJerk-entropy()-Y	tGravityAcc-mean()-X	fBodyBodyAccJerkMag-entropy()	tGravityAcc-min()-X
LAYING	-0.9308	-0.9144	-0.3258	-0.9203	-0.2946
SITTING	-0.9435	-0.9219	0.8804	-0.9304	0.8939
STANDING	-0.9260	-0.8999	0.9330	-0.9147	0.9482
WALKING	0.5361	0.5852	0.9253	0.3886	0.9375
WALKING_DOWNSTAIRS	0.7063	0.6268	0.9165	0.5727	0.9184
WALKING_UPSTAIRS	0.4744	0.4699	0.8554	0.2902	0.8578

The table presents a comparison between the average values of the top ANOVA-ranked features which demonstrate the most strength across various activities. The table helps with interpretation because it connects statistical rankings to actual class behavior while showing which activities produce strong and weak feature response differences.

TABLE IX. TOP FEATURES RANKED BY VARIANCE.

Feature	Variance
fBodyAccJerk-entropy()-X	0.5795
fBodyAccJerk-entropy()-Y	0.5640
tBodyAccJerkMag-entropy()	0.5371
fBodyAcc-entropy()-X	0.5329
tGravityAcc-correlation()-X,Z	0.5221
tGravityAcc-correlation()-Y,Z	0.5039
tGravityAcc-correlation()-X,Y	0.4951

tBodyGyroJerkMag-entropy()	0.4887
tGravityAcc-energy()-X	0.4778
fBodyAccMag-entropy()	0.4754

The table displays the features which demonstrate the greatest total variation between their values. The presence of high variance in data does not ensure that variables will differentiate between classes but it enables researchers to find predictors which contain valuable information for further examination in exploratory data analysis.

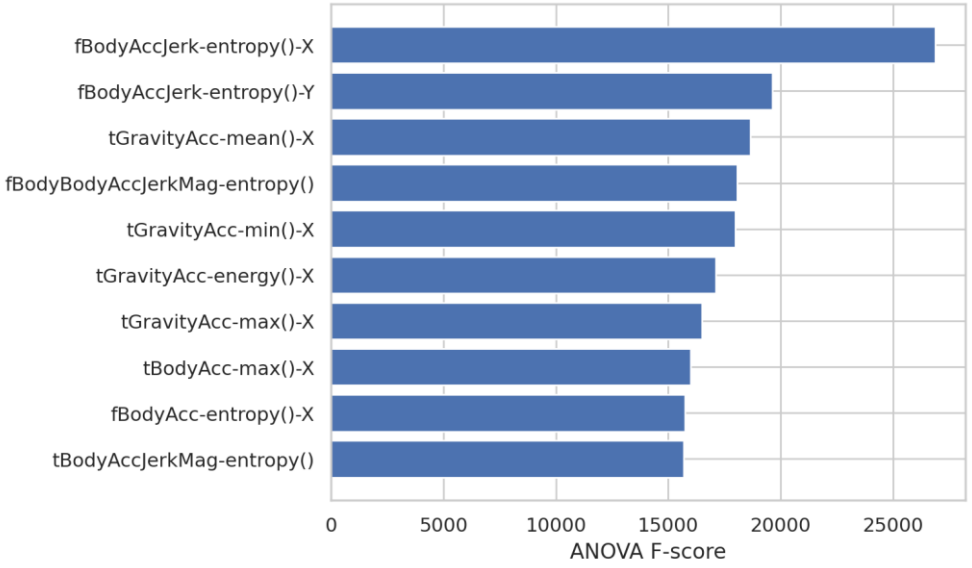

Fig. 6. Top 10 features ranked by ANOVA F-score.

This figure gives a visual ranking of the most discriminative features according to ANOVA. The descending scores make it easy to identify which predictors contribute strongest statistical separation among classes.

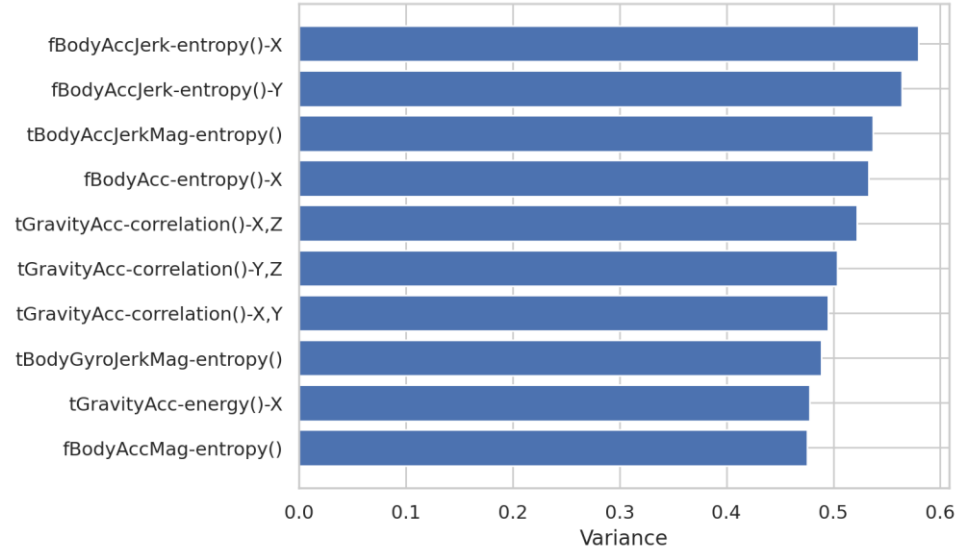

Fig. 7. Top 10 features ranked by variance.

The graph displays the variables which demonstrate the greatest variance in their statistical values. The table presents identical information to Table IX but it shows which features display the most significant value changes throughout the data collection which suggests they might contain useful data that does not relate to specific categories.

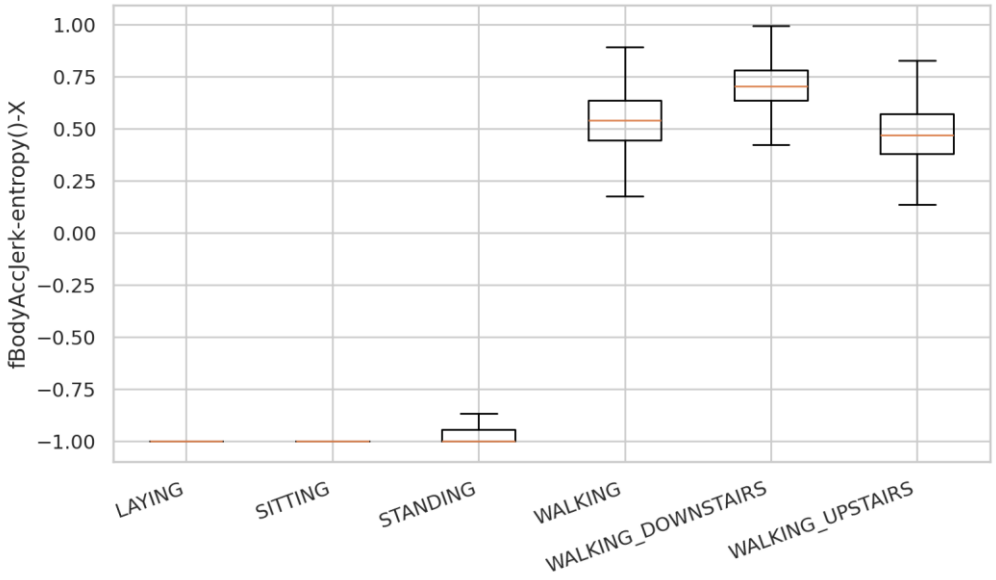

Fig. 8. Distribution of the highest-ranked ANOVA feature across activity classes.

The boxplot in this figure displays the highest ANOVA-ranked feature distribution for every one of the six activities. This method provides exceptional value because it enables users to analyze average values together with distribution patterns and extreme data points through one comprehensive graphical representation.

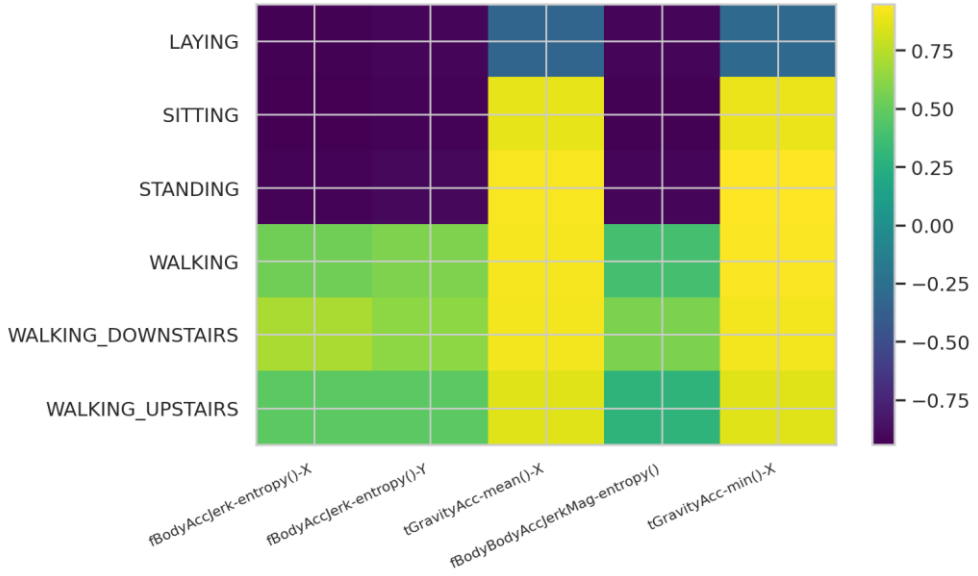

Fig. 9. Heatmap of activity-wise means for the top ANOVA-ranked features.

The illustration shows how the most important features respond on average throughout different activities. The color gradients display unique class patterns which enable users to detect the most powerful features for classes that depend on posture and those that need movement.

F = variance between classes / variance within classes (7)

Equation 7. ANOVA F-Score

The ANOVA F-statistic exists in conceptual form through Equation (7) because it shows the ratio between variance across classes and variance inside classes. A larger value indicates that a feature separates activities more strongly relative to class-internal spread.

3.4 Benchmark Models and Evaluation Metrics

To preserve the lightweight benchmark spirit, four efficient classical models were compared: Logistic Regression, Random Forest, k-Nearest Neighbors, and Gaussian Naive Bayes. The baselines include methods for linear discrimination and ensemble learning and distance-based classification and probabilistic learning. The methods stay relevant because they work with feature-rich sensor descriptors which provide detailed information.

The dataset was split using a stratified 80/20 train-test partition based on the activity label. The evaluation process involved measuring accuracy together with macro-precision and macro-recall and macro F1-score and training time and inference time. The recognition patterns of the best-performing models became clear through two methods which included Random Forest feature importance and confusion analysis. The Macro F1 metric received primary attention because it delivers superior information than accuracy does when dealing with multiple classes which have small differences in their distribution.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (8)$$

Equation 8. Accuracy

Equation (8) defines accuracy as the proportion of correctly classified instances among all evaluated instances. The primary metric provides useful information but requires macro F1 analysis to prevent users from tracking only one performance measure.

$$\text{Macro F1} = (1/C) \Sigma F1\_c \quad (9)$$

Equation 9. Macro F1-Score

Equation (9) defines macro F1-score as the unweighted average of class-level F1 values. The method functions well with multiclass HAR because it assigns identical value to every activity without taking into account the number of times each class appears.

4. RESULTS

4.1 Exploratory Findings

The dataset maintained an acceptable level of balance because LAYING and STANDING samples outnumbered those from WALKING_DOWNSTAIRS and WALKING_UPSTAIRS. The results showed similar levels of contribution from each subject which minimized the possibility that the findings stemmed from a few participants who dominated the study. The data becomes evident through the presentation of Tables II–III and Figures 1–2.

PCA did not fully separate all six classes in two dimensions, which is expected for smartphone HAR. The data showed an organized pattern which clearly distinguished between stationary body positions and moving locomotion patterns. The cumulative variance curve demonstrated that low-dimensional projection enables data inspection yet it fails to substitute the complete feature space for accurate classification systems. The heatmap of correlations verified that several predictors with high variance values maintain strong relationships which supports our future work on feature selection and data compression methods.

4.2 Feature Analysis

Feature ranking revealed that several of the strongest predictors belonged to gravity-related and angle-based groups. The ANOVA table and related figures show that these features carry strong class-separating information, especially for postural activities. The present benchmark achieved its highest feature ranking through a frequency-domain jerk entropy descriptor which Random Forest feature importance demonstrated needed angle(X,gravityMean) and multiple gravity-acceleration descriptors to function properly.

Research findings demonstrate that posture-sensitive orientation information together with dynamic-frequency descriptors improve the recognition system performance. The pattern follows what previous smartphone HAR studies have demonstrated because static pose recognition depends on gravity alignment but dynamic activities require body-motion and jerk descriptors to work effectively [1], [3], [9], [17].

4.3 Lightweight Benchmark Performance

The performance comparison between lightweight benchmark models for smartphone sensor human activity recognition forms the basis of this subsection. The study evaluates each model's ability to classify data while focusing on simple computational requirements which prove useful for systems with limited resources. The models received identical preprocessing treatment and evaluation parameters during their training and testing phases to establish equal evaluation conditions. The evaluation of their performance depends on standard classification metrics which include accuracy and precision and recall and F1-score and on confusion matrix analysis and feature importance interpretation when available. The benchmark seeks to find the best model for accuracy and the most effective lightweight method which delivers optimal performance between prediction power and model transparency and operational speed.

TABLE X. MODEL PERFORMANCE SUMMARY.

Model	Accuracy	Precision (Macro)	Recall (Macro)	F1-score (Macro)	Training Time (s)	Inference Time (s)
Logistic Regression	0.9730	0.9750	0.9754	0.9748	39.1217	0.1065

Random Forest	0.9590	0.9594	0.9598	0.9594	3.3159	0.1648
KNN	0.9560	0.9601	0.9591	0.9593	0.2328	1.2681
Gaussian NB	0.7710	0.7860	0.7728	0.7663	0.2396	0.0820

The main benchmark table of the paper is presented in this table. The table shows the predictive quality and computational practicality of all lightweight models which enables readers to determine the most accurate model along with its performance and efficiency balance.

TABLE XI. CONFUSION MATRIX FOR LOGISTIC REGRESSION.

Class	Pred_LAYING	Pred_SITTING	Pred_STANDING	Pred_WALKING	Pred_WALKING_DOWNSTAIRS	Pred_WALKING_UPSTAIRS
True_LAYING	193	0	0	0	0	0
True_SITTING	0	170	6	0	0	0
True_STANDING	0	19	170	0	0	0
True_WALKING	0	0	0	169	1	0
True_WALKING_DOWNSTAIRS	0	0	0	0	130	0
True_WALKING_UPSTAIRS	0	0	0	0	1	141

The table displays classification results for each class obtained from the best-performing model. The confusion matrix shows the model's steady performance areas and its most common classification errors by displaying detailed information about activity pairs that behave similarly.

TABLE XII. TOP 20 FEATURE IMPORTANCES FROM RANDOM FOREST.

Feature	Importance
angle(X,gravityMean)	0.0388
tGravityAcc-mean()-X	0.0371
tGravityAcc-max()-X	0.0298
tGravityAcc-mean()-Y	0.0294
tGravityAcc-min()-X	0.0284
tGravityAcc-min()-Y	0.0262
tGravityAcc-max()-Y	0.0251
tGravityAcc-energy()-Y	0.0225
angle(Y,gravityMean)	0.0198
tGravityAcc-energy()-X	0.0183
tBodyAccMag-std()	0.0125
fBodyAccJerk-bandsEnergy()-1,16	0.0119
tGravityAccMag-std()	0.0110
fBodyAccJerk-max()-X	0.0104
tBodyAccJerk-std()-X	0.0103

tBodyAccJerk-sma()	0.0102
fBodyAccMag-mean()	0.0098
tGravityAcc-arCoeff()-Y,1	0.0090
fBodyAcc-mad()-X	0.0090
angle(Z,gravityMean)	0.0090

The table displays the top predictive factors which the Random Forest model identifies as most important. The benchmark becomes easier to understand because it reveals which features based on gravity and angle and frequency values most effectively separate different activities.

TABLE XIII. ACCURACY COMPARISON ACROSS BENCHMARKED MODELS.

Model	**Accuracy**
Logistic Regression	0.9730
Random Forest	0.9590
KNN	0.9560
Gaussian NB	0.7710

The table displays accuracy values independently to enable users to determine model rankings without being distracted by additional evaluation metrics. The document aims to make comparisons more understandable instead of introducing new research methods.

TABLE XIV. MACRO F1-SCORE COMPARISON ACROSS BENCHMARKED MODELS.

Model	**F1-score (Macro)**
Logistic Regression	0.9748
Random Forest	0.9594
KNN	0.9593
Gaussian NB	0.7663

The table presents macro F1-scores which deliver better performance evaluation for multiclass classification by assigning equal importance to each class than simple accuracy measures. The system verifies if the highest accuracy level leads to equal performance between all activity categories.

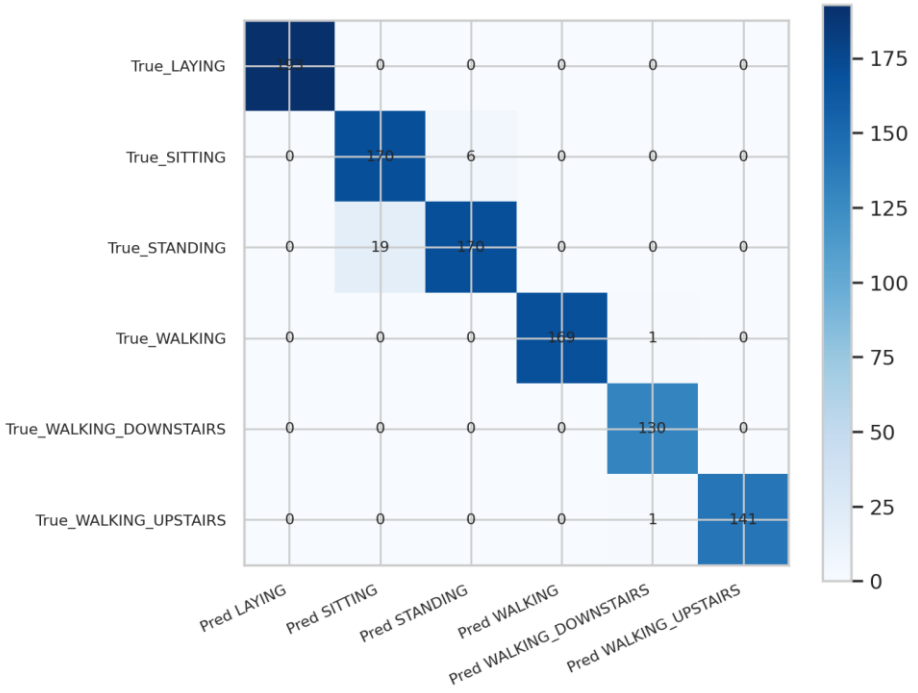

Fig. 10. Confusion heatmap for Logistic Regression.

The heatmap version of the confusion matrix in this figure shows both the correct classifications and the areas where errors tend to cluster. The method allows you to identify the remaining uncertainty which exists between similar postural categories including sitting and standing positions.

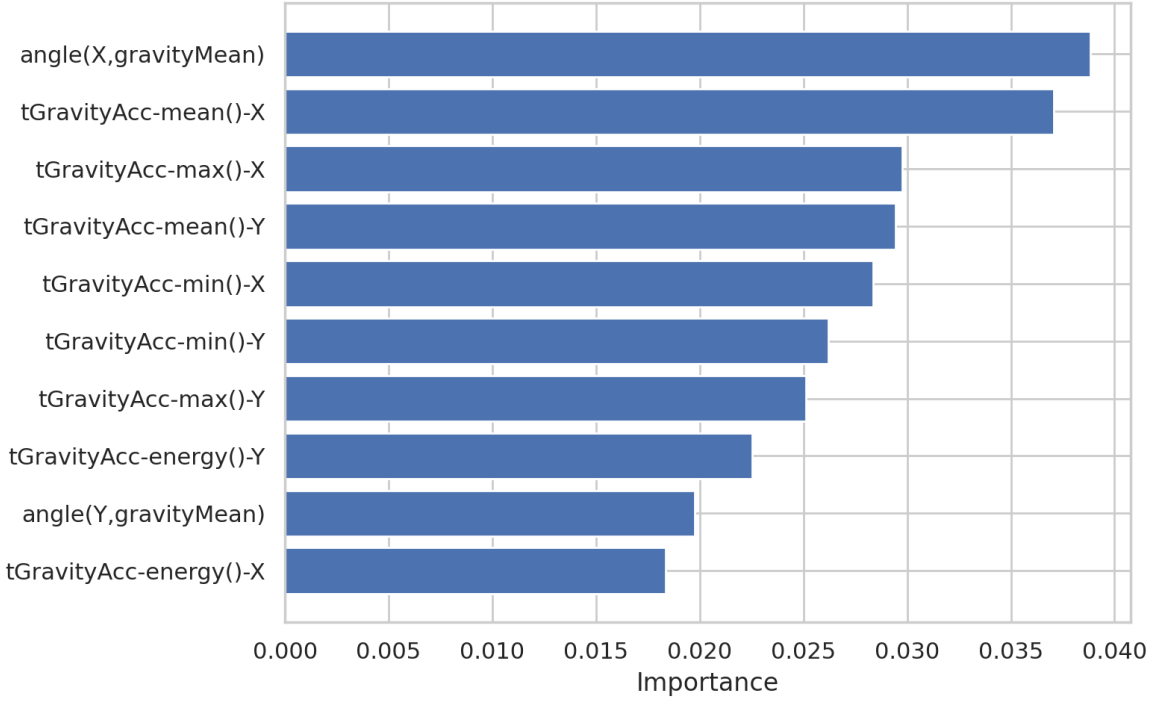

Fig. 11. Top Random Forest feature importances.

The figure displays the feature-importance table through a ranking plot which presents the data in a more user-friendly way. The reader can identify the main variables which control the Random Forest decision system through this visualization while observing the decreasing significance of variables following the highest-ranking predictors.

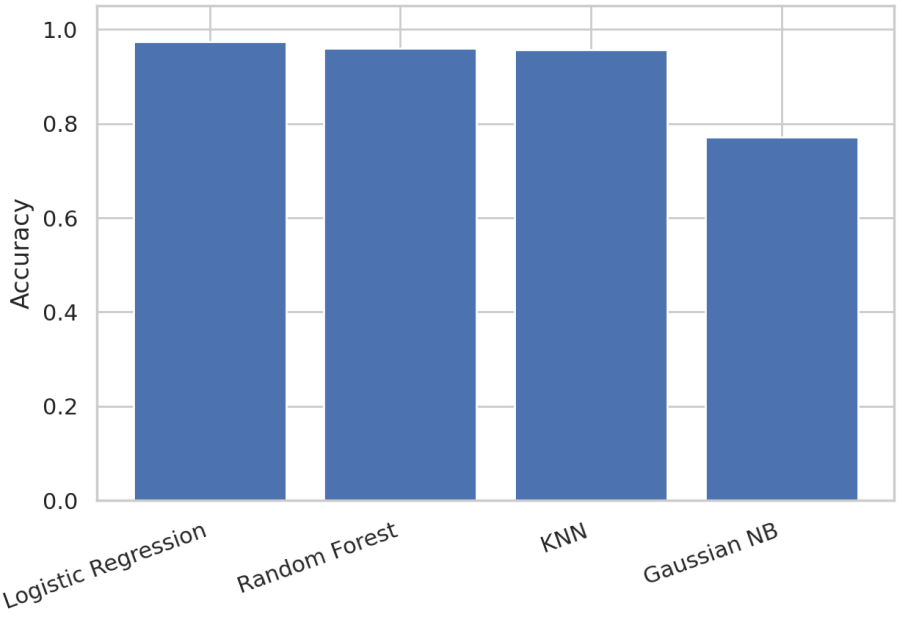

Fig. 12. Accuracy comparison across benchmarked models.

This figure presents model accuracy as a direct side-by-side comparison. It simplifies the interpretation of the benchmark ranking and allows the best-performing model to be identified immediately.

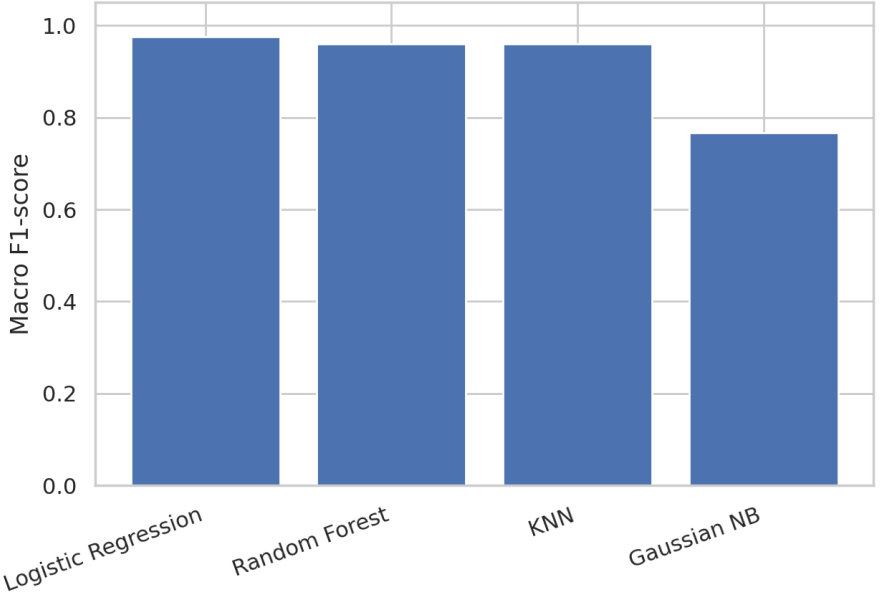

Fig. 13. Macro F1-score comparison across benchmarked models.

This figure visualizes macro F1-scores across models. Because macro F1 emphasizes balanced multiclass performance, it serves as a stronger fairness-oriented complement to the accuracy comparison.

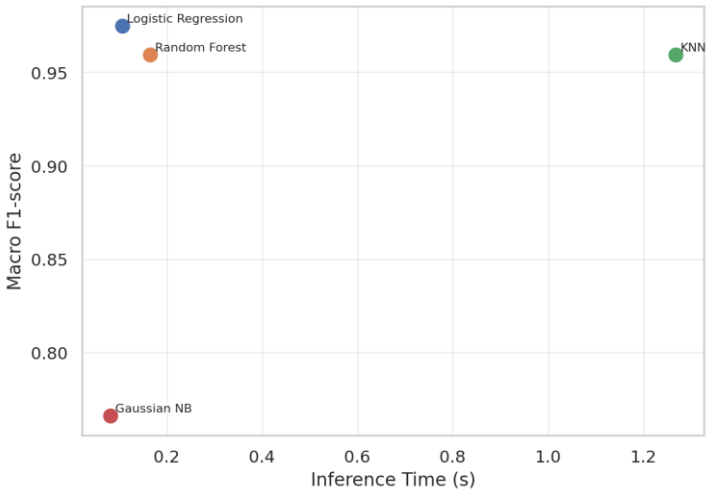

Fig. 14. Inference-time versus macro-F1 trade-off across models.

The diagram demonstrates the core design conflict which lightweight HAR systems encounter because better prediction results should not result in excessive system processing time. The benchmark becomes clear through the plot which shows how well it performs and its readiness for deployment.

Logistic Regression produced the highest performance results by achieving 0.973 accuracy and 0.9750 macro precision and 0.9754 macro recall and 0.9748 macro F1-score. Random Forest achieved an accuracy of 0.959 together with a macro F1-score of 0.9594. KNN obtained a macro F1-score of 0.9593 which matched the other models but its prediction time was significantly longer. The performance of Gaussian Naive Bayes fell short because it achieved 0.771 accuracy and 0.7663 macro F1-score.

The confusion matrix showed that most activities were classified with very high reliability. The test set correctly identified all instances of LAYING while the system performed well in detecting all dynamic locomotion classes. The main error pocket was the sitting-standing boundary, where some STANDING observations were predicted as SITTING and vice versa. The results demonstrate common patterns which seem logical because inertial signatures from nearby static postures show less variation than those observed during locomotion classes according to references [1], [3], [9], and [10].

The trade-off plot shows an important practical result: the best macro F1-score did not require the largest computational burden. The benchmark results showed Logistic Regression delivered the most effective combination of accuracy and operational speed. KNN training occurred quickly but the model needed significant resources for prediction processing. Random Forest performed at a similar level to the linear model yet it failed to produce enough benefit to justify its increased complexity.

5. DISCUSSION

The research findings demonstrate that feature-based lightweight baselines maintain their top performance in smartphone HAR tasks when the feature representation system operates at maximum effectiveness. The recent academic community focuses on compact deep models together with mobile-friendly architectures according to studies [16]–[19], [23] yet practical systems frequently operate without neural networks because their data formats already contain sufficient information. The present benchmark shows that a linear model delivered superior results compared to the more adaptable models. The old tools continue to produce the most powerful strikes at times [24].

Deep learning maintains its complete value regardless of what happens. The document establishes the design space through its presentation. Deep models become most advantageous when raw windows data needs processing and when entire systems require training and when multiple data sources need to be combined [4], [5], [9]–[11]. The combined evaluation of transparency and training simplicity and inference footprint and interpretability shows classical baselines remain superior to other methods when working with existing engineered descriptors.

The features which base their calculations on gravity and angles match the way people perform their daily tasks during smartphone HAR operations. Static postures show different device orientations against gravity but walking behaviors display patterned movement through jerk and frequency-domain measurements. The optimal deployment-oriented feature set needs to include both static and dynamic descriptors from multiple signal families instead of requiring a large number of features.

The benchmark also reveals where classical approaches struggle. Sitting and standing remained the hardest pair to separate. Research has demonstrated through multiple studies that postural categories tend to intersect because of how people position their phones and how they stand and move their bodies [1], [3], [13], [15]. The subject-independent and position-independent experimental conditions would probably cause these errors to increase in number. The benchmark requires

future research to develop subject-independent evaluation methods together with compact deep baselines and feature-compression assessment techniques.

The research needs to disclose its existing restrictions in a straightforward manner. The benchmark operates with one uploaded dataset instead of utilizing various publicly accessible datasets. The evaluation process uses label-based stratification instead of achieving complete independence from subject data. Third, no raw-signal deep model was trained in the present manuscript. The research results remain valid despite these limitations but they set limits on how strongly we can prove our hypotheses. The paper presents a lightweight benchmark through its feature-based approach which does not function as a comprehensive leaderboard document.

6. CONCLUSION

The research paper delivers a complete journal-style article which employs smartphone sensors to identify human activities through a dataset containing refined features and an efficient benchmarking system. The study developed a single operational system which combined data cleaning methods with descriptive and discriminative analysis and dimensionality reduction and feature ranking and classical model benchmarking and performance interpretation tools.

The research data demonstrated that Logistic Regression produced the highest complete performance by achieving 0.973 accuracy and 0.9748 macro F1-score and it provides efficient results during inference operations. The most useful information came from Gravity-related and angle-based variables yet the system faced its most significant challenge in distinguishing between sitting and standing positions. These observations are consistent with broader HAR literature and reinforce the practical relevance of well-engineered feature spaces.

The broader lesson is straightforward: smartphone HAR does not always require heavyweight modeling to achieve strong results. When the feature representation reaches its full development stage and deployment restrictions exist in reality, standardized data transformation methods together with basic classification algorithms will generate top-tier results. The benchmark needs future development to support subject-independent validation and compact neural baselines and enhanced feature compression methods for achieving authentic device-based deployment standards.

Funding:

The authors declare that no specific financial aid or sponsorship was received from governmental, private, or commercial entities to support this study. The research was solely financed by the authors' own contributions.

Conflicts of Interest:

The authors declare that there are no conflicts of interest in this study.

Acknowledgment:

The authors express their heartfelt appreciation to their institutions for the essential support and motivation provided throughout the research period.

References

[1] V. Dentamaro, V. Gattulli, D. Impedovo, and F. Manca, “Human activity recognition with smartphone-integrated sensors: A survey,” *Expert Systems with Applications*, vol. 246, Art. no. 123143, 2024.

[2] M. Straczkiewicz, P. James, and J.-P. Onnela, “A systematic review of smartphone-based human activity recognition methods for health research,” *npj Digital Medicine*, vol. 4, no. 1, Art. no. 148, 2021.

[3] J. Morales and D. Akopian, “Physical activity recognition by smartphones: A survey,” *Biocybernetics and Biomedical Engineering*, vol. 37, no. 3, pp. 388–400, 2017.

[4] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, “Deep learning for sensor-based activity recognition: A survey,” *Pattern Recognition Letters*, vol. 119, pp. 3–11, 2019.

[5] H. F. Nweke, Y. W. Teh, M. A. Al-Garadi, and U. R. Alo, “Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges,” *Expert Systems with Applications*, vol. 105, pp. 233–261, 2018.

[6] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, “A public domain dataset for human activity recognition using smartphones,” in *Proc. European Symposium on Artificial Neural Networks (ESANN)*, Bruges, Belgium, 2013, pp. 437–442.

[7] J. L. Reyes-Ortiz, D. Anguita, A. Ghio, L. Oneto, and X. Parra, *Human Activity Recognition Using Smartphones* [Dataset]. Irvine, CA, USA: UCI Machine Learning Repository, 2013.

[8] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, “Activity recognition using cell phone accelerometers,” *ACM SIGKDD Explorations Newsletter*, vol. 12, no. 2, pp. 74–82, 2011.

[9] C. A. Ronao and S.-B. Cho, “Human activity recognition with smartphone sensors using deep learning neural networks,” *Expert Systems with Applications*, vol. 59, pp. 235–244, 2016.

[10] F. J. Ordóñez and D. Roggen, “Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition,” *Sensors*, vol. 16, no. 1, Art. no. 115, 2016.

[11] E. Ramanujam, T. Perumal, and S. Padmavathi, "Human activity recognition with smartphone and wearable sensors using deep learning techniques: A review," *IEEE Sensors Journal*, vol. 21, no. 12, pp. 13029–13040, 2021.

[12] L. M. Dang, M. Piran, D. Han, et al., "Sensor-based and vision-based human activity recognition: A comprehensive survey," *Pattern Recognition*, vol. 108, Art. no. 107561, 2020.

[13] A. Ferrari, D. Micucci, M. Mobilio, and P. Napoletano, "Trends in human activity recognition using smartphones," *Journal of Reliable Intelligent Environments*, vol. 7, no. 3, pp. 189–213, 2021.

[14] D. Micucci, M. Mobilio, and P. Napoletano, "UniMiB SHAR: A dataset for human activity recognition using acceleration data from smartphones," *Applied Sciences*, vol. 7, no. 10, Art. no. 1101, 2017.

[15] D. Garcia-Gonzalez, M. Rivero, and J. R. Rabuñal, "A public domain dataset for real-life human activity recognition using smartphone sensors," *Sensors*, vol. 20, no. 8, Art. no. 2200, 2020.

[16] S. R. Sekaran, P. Y. Han, and O. S. Yin, "Smartphone-based human activity recognition using lightweight multiheaded temporal convolutional network," *Expert Systems with Applications*, vol. 227, Art. no. 120132, 2023.

[17] S. AlMuhaideb, L. AlAbdulkarim, D. M. AlShahrani, H. AlDhubaib, and D. E. AlSadoun, "Achieving more with less: A lightweight deep learning solution for advanced human activity recognition," *Sensors*, vol. 24, no. 16, Art. no. 5436, 2024.

[18] X. Gong, X. Zhang, and N. Li, "Lightweight human activity recognition method based on the MobileHARC model," *Systems Science & Control Engineering*, vol. 12, no. 1, Art. no. 2328549, 2024.

[19] P. Agarwal and M. Alam, "A lightweight deep learning model for human activity recognition on edge devices," *Procedia Computer Science*, vol. 167, pp. 2364–2373, 2020.

[20] G. M. Weiss, *WISDM Smartphone and Smartwatch Activity and Biometrics Dataset* [Dataset]. Irvine, CA, USA: UCI Machine Learning Repository, 2019.

[21] M. Alanazi, A. Aljallal, and A. Alotaibi, "Human activity recognition through smartphone inertial sensors using machine learning and deep learning approaches," *Engineering, Technology & Applied Science Research*, vol. 14, no. 2, pp. 13359–13364, 2024.

[22] L. A. S. Zendron, F. C. Delicato, and D. G. Gomes, "Enhancing human activity recognition with machine learning: Insights from smartphone accelerometer and magnetometer data," *PeerJ Computer Science*, vol. 11, Art. no. e3137, 2025.

[23] P. Wang, F. Guo, F. Gu, M. Li, and X. Long, "MobileHAR: A lightweight and efficient human activity recognition model based on inverted residual inception block," in *Proc. 20th Int. Conf. Mobility, Sensing and Networking (MSN)*, 2024, pp. 834–841.

[24] A. Ankita, S. Gupta, and S. Kumar, "An efficient and lightweight deep learning model for human activity recognition," *Sensors*, vol. 21, no. 11, Art. no. 3845, 2021.