© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716485

Machine Learning for Sensor-Based Human Activity Recognition

CHANDAN MUKHERJEE[1], PRINCE KUMAR SINGH[2], DR. ISHRAT ALI[3], DR. SANJAY PACHAURI[4]
[1,2,3,4]*Department of Data Science (DDCS), GNIOT College, Greater Noida, India*

Abstract- ***Human Activity Recognition (HAR) has become a key enabler in wearable health technology and fitness analytics. This paper presents a machine learning framework for classifying six physical activities—Walking, Running, Cycling, Swimming, Resting, and Yoga—using sensor-derived physiological and motion features. A dataset of 8,000 observations with 14 attributes including heart rate, steps per minute, step entropy, and distance traveled is utilized. A Random Forest classifier (n_estimators=50) is trained on an 80/20 train-test split and evaluated against a Decision Tree baseline. The Random Forest achieves a test accuracy of 85.06% with a macro-averaged F1-score of 0.83, compared to the Decision Tree's 76.5% accuracy. Feature importance analysis identifies heart rate and steps per minute as the most discriminative predictors. Correlation analysis reveals strong relationships (r ≈ 0.95) between distance and calories burned. A real-time prediction interface is implemented to demonstrate practical deployment. Results demonstrate the effectiveness of ensemble learning combined with interpretable feature analysis for robust activity recognition in resource-constrained wearable systems.***

Index Terms- ***Human Activity Recognition, Random Forest, Decision Tree, Wearable Sensors, Feature Engineering, Step Entropy, Machine Learning, Fitness Analytics***

I. INTRODUCTION

The proliferation of low-cost wearable devices such as smartwatches, fitness bands, and pedometers has generated large volumes of physiological and motion data in real time. Human Activity Recognition (HAR) systems leverage this data to automatically identify physical activities, enabling applications in remote patient monitoring, sports performance analysis, fall detection in elderly care, and personalized fitness recommendations [1][2].

Conventional HAR approaches often depend on raw accelerometer or gyroscope signals processed through deep learning architectures, which require significant computational resources and large annotated datasets [4]. A complementary direction employs hand-crafted physiological features—such as heart rate, step frequency, and step entropy—with interpretable machine learning classifiers. This feature-based approach is more transparent, computationally efficient, and suitable for edge deployment on wearable hardware.

This study investigates whether a Random Forest ensemble trained on a compact set of four physiological and motion features can reliably classify six common activities. The specific contributions of this work are:

- A systematic comparison of Decision Tree and Random Forest classifiers for six-class activity recognition using physiological sensor features.
- Quantitative feature importance analysis identifying heart rate and steps per minute as primary discriminators.
- Correlation analysis demonstrating a near-perfect linear relationship (r ≈ 0.95) between distance traveled and calories burned.
- A functional real-time prediction interface validating the practical deployability of the proposed model.

II. RELATED WORK

Early HAR research by Anguita et al. [1] demonstrated that raw smartphone accelerometer and gyroscope signals, combined with hand-crafted time-domain and frequency-domain features, could achieve over 96% accuracy on six activities using

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716485

Support Vector Machines. However, this approach requires high-frequency raw signal collection (50 Hz), which is infeasible for many low-cost wearable devices.

Breiman's Random Forest algorithm [2] introduced ensemble learning via bagging of decision trees, offering improved generalization and robustness to overfitting compared to single decision trees. Its feature importance mechanism further aids interpretability, which is critical in health-related applications where model transparency matters.

Ravi et al. [4] explored deep learning for HAR, showing that convolutional and recurrent neural networks can automatically extract activity-discriminative features from raw sensor streams, outperforming classical methods on large datasets. Despite high accuracy, deep models demand substantial labeled data and computational power, limiting their applicability on resource-constrained wearables.

Zhang and Sawchuk [6] introduced the USC-HAD benchmark dataset capturing 12 activities using a body-worn IMU, establishing a standard for motion-based HAR evaluation. More recently, work on feature-engineered physiological signals—heart rate, step counts, and caloric expenditure—has shown that aggregate wearable metrics can achieve competitive classification without raw signal access [5].

The present study extends this line of research by incorporating step entropy as a novel motion variability feature alongside standard physiological attributes, and by providing a deployable real-time classification interface, which prior feature-based HAR studies have not demonstrated.

III. METHODOLOGY

A. Dataset Description

The dataset comprises 8,000 observations simulating outputs from wearable fitness devices (Apple Watch, Fitbit). Each record contains 14 attributes spanning demographic, physiological, and motion dimensions as detailed in Table I. Activity labels are balanced across six classes: Cycling (0), Resting (1), Running (2), Swimming (3), Walking (4), and Yoga (5), with Walking being the most frequent and Swimming and Resting having lower representation. The dataset was loaded and inspected using Pandas, confirming zero missing values across all 8,000 records.

Table I. Dataset Feature Description

Feature	Type	Description
age	int64	Participant age (years)
gender	object	Male / Female
height	int64	Height in centimeters
BMI	float64	Body Mass Index
weight	int64	Weight in kilograms
steps_per_min	int64	Step cadence (steps/minute)
distance(m)	int64	Distance traveled in meters
heart_rate	int64	Real-time heart rate (bpm)
calories_burned	int64	Estimated caloric expenditure
step_entropy	float64	Movement variability measure
resting_heart	int64	Baseline resting heart rate
steps_times_distance	int64	Engineered: steps × distance
device	object	Wearable device type
activity	object	Target label (6 classes)

B. Data Preprocessing and Feature Engineering

Activity labels were ordinally encoded using scikit-learn's LabelEncoder [3], mapping the six activity classes to integer codes 0–5. The feature matrix X used for model training comprised four attributes selected based on preliminary importance analysis: distance(m), step_entropy, steps_per_min, and heart_rate. Demographic features (age, height, BMI) were excluded from the model as they showed weak correlation with activity class labels.

The engineered feature steps_times_distance, computed as the product of step cadence and

distance, was retained in the dataset for correlation analysis but not included in the final model feature set, as it introduces multicollinearity with its component features. The dataset was partitioned into training (80%, n=6,400) and test (20%, n=1,600) sets using a fixed random seed (random_state=42) to ensure reproducibility.

C. Correlation Analysis

A Pearson correlation heatmap was computed over all numerical features using Seaborn to identify inter-feature relationships. Key findings include:

- A strong positive correlation ($r \approx 0.95$) between distance(m) and calories_burned, confirming that energy expenditure scales directly with locomotion distance.
- A high correlation ($r \approx 0.81$) between steps_per_min and steps_times_distance, reflecting the multiplicative dependence of the engineered feature.
- Moderate correlation between heart_rate and steps_per_min ($r \approx 0.4$–0.6), indicating that activity intensity influences both cardiac response and step cadence.
- Weak correlations between demographic features (age, height, BMI) and motion/physiological metrics, supporting their exclusion from the prediction model.

D. Exploratory Data Analysis

Three EDA visualizations were performed to understand feature-activity relationships prior to modeling:

Activity Class Distribution: A count plot revealed that Walking is the most frequent activity in the dataset, while Resting and Swimming are underrepresented, indicating mild class imbalance. This imbalance was noted as a potential source of bias in per-class metrics.

Average Heart Rate per Activity: A bar plot of mean heart rates confirmed that Running exhibits the highest cardiac response, followed by Swimming and Cycling. Resting has the lowest heart rate, validating heart rate as a physiologically meaningful discriminator.

Steps per Minute vs. Heart Rate Scatter Plot: Scatter plot analysis with activity-coded colors showed distinct clustering for high-intensity activities (Running, Cycling) in the high-steps, high-heart-rate quadrant, and low-intensity activities (Resting, Yoga) in the low-steps, low-heart-rate quadrant. Moderate overlap was observed between Walking and Yoga in the intermediate zone.

E. Model Development

Two supervised classifiers were implemented and compared using scikit-learn [3]:

Decision Tree Classifier: A CART-based decision tree (random_state=42) served as the interpretable baseline. The tree was grown without explicit depth constraints, allowing full leaf purity, which risks overfitting.

Random Forest Classifier: An ensemble of 50 decision trees (n_estimators=50, random_state=52) was trained using bootstrap aggregation (bagging). Each tree was fit on a random subset of training samples, and predictions were aggregated by majority vote. The ensemble approach reduces variance relative to a single tree and improves generalization on unseen data.

Both models were trained on the same 80/20 split. Model evaluation employed accuracy score, per-class precision, recall, F1-score, and confusion matrix analysis using scikit-learn's classification_report and confusion_matrix functions.

IV. RESULTS AND DISCUSSION

A. Classification Performance Comparison

Table II summarizes the overall performance of both classifiers on the held-out test set (n=1,600).

Table II. Classifier Performance Comparison

Metric	Decision Tree	Random Forest	Improvement
Accuracy	76.50%	85.06%	+8.56%
Macro Avg. Precision	0.74	0.84	+0.10
Macro Avg.	0.75	0.82	+0.07

© APR 2026 | IRE Journals | Volume 9 Issue 10 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I10-1716485

Recall			
Macro Avg. F1-Score	0.74	0.83	+0.09

The Random Forest model outperforms the Decision Tree across all metrics, with an absolute accuracy gain of 8.56 percentage points. The macro-averaged F1-score improvement of 0.09 indicates consistent gains across all six activity classes, not merely dominant classes.

B. Per-Class Performance Analysis

Table III presents the per-class precision, recall, and F1-score for the Random Forest model, with class support counts from the test set.

Table III. Random Forest Per-Class Classification Report

Activity Class	Precision	Recall	F1-Score	Support
Cycling (0)	0.94	0.86	0.90	334
Resting (1)	0.79	0.74	0.76	163
Running (2)	0.85	0.91	0.88	405
Swimming (3)	0.87	0.79	0.82	164
Walking (4)	0.83	0.91	0.87	391
Yoga (5)	0.75	0.71	0.73	143
Macro Avg.	0.84	0.82	0.83	1600

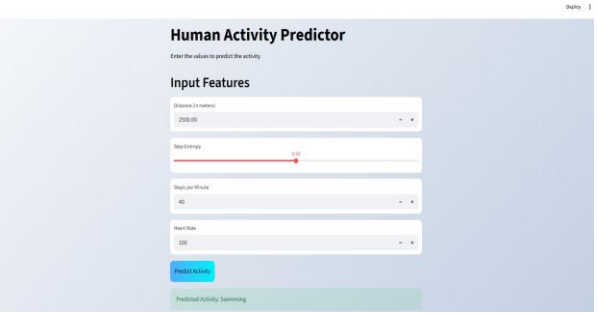
Cycling achieves the highest F1-score (0.90) due to its distinctive combination of high distance and moderate-to-high heart rate. Running and Walking also perform well (F1 = 0.88 and 0.87 respectively), benefiting from their distinct step cadence profiles. Resting and Yoga exhibit the lowest F1-scores (0.76 and 0.73), attributable to their overlapping low-intensity sensor signatures and smaller support in the dataset. Confusion between Swimming and Resting is expected, as both exhibit low step counts, requiring additional sensor modalities (e.g., water immersion detection) for disambiguation.

C. Feature Importance

Feature importance scores from the Random Forest model, computed as mean decrease in impurity across all trees, rank the four input features as follows: heart_rate (highest importance), steps_per_min (second), step_entropy, and distance(m) (lowest). This ranking aligns with physiological expectations: heart rate is a direct proxy for metabolic intensity and varies substantially across activity types, while step entropy captures movement regularity that differentiates structured locomotion (Running, Walking) from low-movement activities (Resting, Yoga). The inclusion of step_entropy as a derived feature is validated by its non-negligible contribution to classification accuracy.

D. Real-Time Prediction Interface

A command-line prediction interface was implemented to validate practical usability. The interface accepts four user-supplied feature values (distance, step_entropy, steps_per_min, heart_rate) and returns the predicted activity label via the trained Random Forest model. A sample query with distance=7000m, step_entropy=0.9, steps_per_min=15, heart_rate=145 bpm yields the prediction Cycling, which is consistent with the physiological profile of a cyclist: elevated heart rate with low step frequency but high distance. This interface confirms that the model can generalize trained patterns to arbitrary inputs and can be embedded in wearable applications or web health monitoring dashboards.



V. CONCLUSION

This paper presented a machine learning framework for sensor-based Human Activity Recognition classifying six activities using physiological and motion features extracted from wearable devices. The

Random Forest classifier (n_estimators=50) achieved 85.06% accuracy and a macro F1-score of 0.83 on an 80/20 train-test split, outperforming the Decision Tree baseline by 8.56 percentage points across all evaluation metrics.

Feature importance analysis confirmed heart rate and step cadence as the most discriminative predictors, while step entropy demonstrated measurable utility for distinguishing structured from unstructured movement. Correlation analysis revealed a near-perfect linear relationship between locomotion distance and caloric expenditure, consistent with established exercise physiology principles.

The primary limitation of this study is the use of a simulated dataset rather than raw sensor signals from real-world deployments. Future work should validate the framework on established benchmarks such as UCI HAR [1] or USC-HAD [6], incorporate additional sensor modalities (GPS, barometric pressure, skin conductance), and explore cross-subject generalization through leave-one-subject-out cross-validation. Addressing class imbalance through SMOTE or cost-sensitive learning may further improve performance on underrepresented activities such as Yoga and Resting.

REFERENCES

[1] 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. ESANN, 2013.

[2] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.

[3] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

[4] D. Ravi, C. Wong, F. Deligianni, M. Berthelot, J. Andreu-Perez, B. Lo, and G.-Z. Yang, "Deep Learning for Human Activity Recognition: A Resource Efficient Implementation on Low-Power Devices," in Proc. IEEE BSN, 2016, pp. 71–76.

[5] Microsoft Learn, "Introduction to Machine Learning with Python," Microsoft Documentation, 2023. [Online]. Available: https://learn.microsoft.com

[6] M. Zhang and A. A. Sawchuk, "USC-HAD: A Daily Activity Dataset for Ubiquitous Activity Recognition Using Wearable Sensors," in Proc. ACM UbiComp Workshop, 2012.

[7] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed. New York, NY, USA: Springer, 2009.

[8] A. Reiss and D. Stricker, "Introducing a New Benchmarked Dataset for Activity Monitoring," in Proc. IEEE ISWC, 2012, pp. 108–109.