International Journal of Computer Techniques–IJCT Volume 13 Issue 2, March-April - 2026

Open Access and Peer Review Journal ISSN 2394-2231 https://ijctjournal.org/

RESEARCH ARTICLE OPEN ACCESS

A Comparative Machine Learning Approach for Smartphone-Based Human Activity Recognition Using Feature Engineering

Aditya Pratap Yadav, Meidingu Lourembam,
Suhas Achappa, Achyut N Gowda

Under the guidance of
Prasanth T
Associate Professor, CSE
Reva University, Bengaluru

Computer Science and Engineering, Reva University, and Bengaluru

Email : Meidinglourem@gma il.com

Email: yadav153adity@gma il.com

Abstract:

This Abstract Human Activity Recognition (HAR) has become an essential study area in applications such as medical care observation, fitness monitoring, and smart environments. With the rising availability of smartphone sensors, scalable and real-time activity classification has become practical and workable. This paper presents a comparative analysis of four supervised machine learning models; k-Nearest Neighbors (k-NN), Logistic Regression, Random Forest; and Support Vector Machine (SVM), using the UCI HAR dataset. The models are assessed applying cross-validation and numerous and manifold performance metrics. Experimental outcomes show that Random Forest achieves the supreme and paramount accuracy of 93.07% with robust generalization performance. Feature significance analysis discloses that gravity-based and angle-related features considerably affect classification. The study highlights the effectiveness of ensemble approaches for feature-based HAR systems and offers insights for real-world deployment.

Keywords — HAR, Machine Learning, Random Forest, Smartphone Sensors, Classification

I. INTRODUCTION

Human Activity Recognition (HAR) includes the automatic identification of physical activities performed by individuals utilizing sensor data. The pervasive and far-reaching adoption of smartphones equipped with motion sensors such as accelerometers and gyroscopes has made it possible to accumulate real-time activity data efficiently. This has enabled the development of intelligent systems for applications incorporating medical care tracking, rehabilitation, and smart living environments. Traditional rule-based methods for activity recognition frequently lack adaptability and fail to generalize across varied datasets. In contrast,

machine learning methods provide a data-driven solution by learning patterns directly from sensor data. However, selecting an apt model that balances accuracy, computational productivity, and interpretability remains a notable difficulty. This study addresses this difficulty by comparing various machine learning algorithms and identifying the most effective model for feature-based HAR.

II. RELATED WORK

A Several studies have explored the application of machine learning techniques in HAR. Early research focused on feature engineering combined with classifiers such as Decision Trees, k-NN, and Support Vector Machines, demonstrating reliable performance on structured datasets. The introduction of the UCI HAR dataset further standardized evaluation and enabled comparative studies across different models.

Recent advancements have shifted toward deep learning techniques such as Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks, which automatically extract features from raw sensor data. Although these methods achieve higher accuracy, they require significant computational resources and are less interpretable. In contrast, classical machine learning models remain relevant due to their efficiency and suitability for real-time mobile deployment.

III. DATASET AND FEATURE REPRESENTATION

The dataset used in this study is the UCI HAR dataset, which contains 10,299 observations with 561 engineered features and six activity classes. Each observation corresponds to a fixed-width sliding window of 2.56 seconds, with sensor signals sampled at a frequency of 50 Hz.

The feature set includes both time-domain and frequency-domain characteristics derived from raw sensor data. Time-domain features capture statistical properties such as mean and standard deviation, while frequency-domain features are obtained using Fast Fourier Transform (FFT). Additional features such as signal magnitude area,

IV. METHODOLOGY

The HAR problem is formulated as a supervised classification task, where the goal is to learn a function that maps feature vectors to activity labels. Four machine learning models are implemented and evaluated in this study.

Logistic Regression serves as a baseline model due to its simplicity and interpretability. The k-NN algorithm classifies data based on similarity measures in the feature space. Support Vector Machine (SVM) constructs an optimal hyperplane to separate classes, utilizing an RBF kernel to handle nonlinear patterns. Random Forest, an ensemble learning method, combines multiple decision trees to improve classification accuracy and reduce overfitting.

V. EXPERIMENTAL SETUP

The dataset is divided into training and testing sets using a 70–30 split. To ensure robustness, 5-fold cross-validation is applied. Feature standardization is performed to normalize the data before training.

Model performance is evaluated using accuracy, precision, recall, and F1-score. These metrics provide a comprehensive evaluation of classification effectiveness across different activity classes.

VI. RESULTS AND DISCUSSION

The experimental results demonstrate that Random Forest achieves the highest accuracy of 93.07%, outperforming all other models. SVM shows competitive performance, while Logistic Regression and k-NN yield comparatively lower accuracy.

Cross-validation results indicate that Random Forest exhibits the lowest variance, highlighting its strong generalization capability. The confusion matrix analysis reveals that most misclassifications occur between activities such as sitting and standing, which share similar posture characteristics.

Confusion Matrix Analysis:
The confusion matrix of the Random Forest model is presented in Fig. 1. It can be observed that the model correctly classifies most activity classes with high accuracy. Minor misclassifications occur between activities such as sitting and standing, which exhibit similar posture characteristics. This indicates that while the model performs well overall, distinguishing between closely related static activities remains challenging.

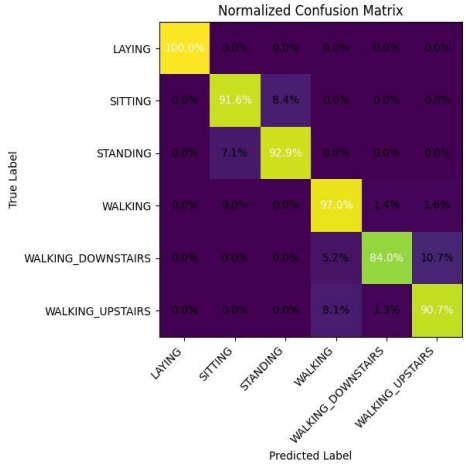

Fig. 1. Confusion Matrix of Random Forest Model

VII. FEATURE IMPORTANCE ANALYSIS

Feature importance analysis indicates that gravity-related features and angle-based features contribute significantly to classification performance. These features are particularly effective in distinguishing static activities. Dynamic activities such as walking rely more on frequency-domain features and signal energy, highlighting the importance of combining multiple feature types. The feature importance distribution is illustrated in Fig. 2, highlighting the most influential features contributing to classification performance.

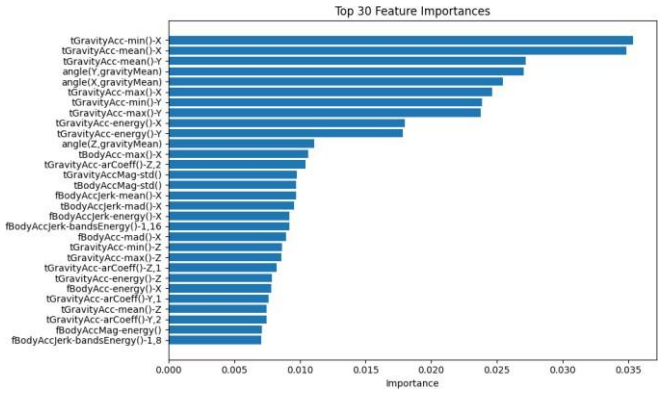

Fig. 2. Feature Importance of Top Features

VIII. STATISTICAL VALIDATION

A paired t-test is conducted to compare the performance of SVM and Random Forest. The obtained p-value is less than 0.05, indicating that the improvement achieved by Random Forest is statistically significant and not due to random variation.

IX. CONCLUSION

This paper presents a comparative analysis of four machine learning models for smartphone-based Human Activity Recognition. The results demonstrate that Random Forest provides the best performance in terms of accuracy and stability. Its ability to handle high-dimensional data and reduce overfitting makes it a suitable choice for real-world applications.

Future work will focus on exploring deep learning approaches, real-time mobile deployment, and personalized activity recognition systems

ACKNOWLEDGMENT

The author would like to express sincere gratitude to the School of Computer Science and Engineering, REVA University, for providing the necessary resources and support to conduct this research. Special thanks are extended to the faculty members and project mentors for their valuable guidance, constructive feedback, and continuous encouragement throughout the study.

REFERENCES

[1] D. Anguita et al., "A Public Domain Dataset for Human Activity Recognition Using Smartphones," ESANN, 2013.

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

[3] T. Cover and P. Hart, "Nearest Neighbor Pattern Classification," IEEE Transactions on Information Theory, 1967.

[4] C. Cortes and V. Vapnik, "Support Vector Networks," Machine Learning, 1995.

[5] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, 2011.

[6] J. R. Kwapisz et al., "Activity Recognition using Cell Phone Accelerometers," ACM SIGKDD, 2011.