Unsupervised Analysis of Lifestyle and Physical Activity Patterns Using Wearable Sensor Data

Carlxen Brieyl Duran
College of Computing and Information Technologies
National University
Manila, Philippines
durancp@students.national-u.edu.ph

Mark Rid Ramirez
College of Computing and Information Technologies
National University
Manila, Philippines
ramirezmb@students.national-u.edu.ph

***Abstract*—This study investigates latent lifestyle and physical activity patterns using wearable sensor data and unsupervised machine learning techniques. Daily Fitbit activity summaries from 35 participants (457 observations) were analyzed through a structured pipeline including exploratory data analysis, logarithmic transformation, z-score normalization, and Principal Component Analysis (PCA). Four clustering algorithms—K-Means, Hierarchical Clustering, DBSCAN, and Gaussian Mixture Models (GMM)—were systematically compared using Silhouette Score, Calinski–Harabasz Index, and Davies–Bouldin Index. Three distinct behavioral clusters emerged: (1) sedentary individuals characterized by low step counts and prolonged inactivity, (2) highly active individuals exhibiting elevated movement intensity and calorie expenditure, and (3) moderately active individuals demonstrating mixed movement patterns with substantial sedentary duration. Among the evaluated models, K-Means achieved the strongest overall clustering performance, indicating superior compactness and separation of behavioral groups. The findings demonstrate the effectiveness of combining wearable sensor data with unsupervised learning to uncover interpretable lifestyle typologies, with implications for personalized health interventions and data-driven physical activity monitoring.**

***Index Terms*—Wearable sensor data, physical activity clustering, K-Means, Principal Component Analysis (PCA), unsupervised learning, lifestyle segmentation, sedentary behavior.**

I. Introduction

Wearable sensor technologies such as accelerometers and fitness trackers enable continuous monitoring of physical activity, step counts, sedentary behavior, and lifestyle patterns. These devices generate objective, high-resolution behavioral data that overcome limitations of self-reported measures [1], [2], [3]. With increasing adoption of wearable devices, large-scale activity datasets are now available for data-driven analysis. Physical activity is strongly associated with improved health outcomes. The World Health Organization recommends regular moderate-to-vigorous activity and reduced sedentary time to prevent chronic diseases [4]. Empirical studies published in JAMA [5] and The Lancet Public Health [6] confirm that higher daily step counts are associated with lower mortality risk and improved health outcomes. However, individuals accumulate activity differently across the day, resulting in heterogeneous 24-hour movement patterns [7]. Unsupervised machine learning provides a powerful framework for identifying latent behavioral patterns without predefined labels. Clustering techniques such as K-Means, hierarchical clustering, DBSCAN, and Gaussian Mixture Models (GMM) allow the discovery of natural groupings within wearable sensor data. This study applies and compares these algorithms to analyze lifestyle and physical activity patterns derived from wearable devices.

II. Literature Review

Recent research highlights the importance of clustering approaches in understanding physical activity behavior. Nawrin et al. [7] demonstrated that machine learning clustering reveals diverse 24-hour step-counting patterns, emphasizing variability in daily movement structures. Similarly, Pontin et al. [1] used unsupervised learning to characterize temporal step-count behavior from smartphone data.

Hierarchical clustering has been used to identify meaningful activity profiles. Shim et al. [8] showed that wearable-derived accelerometer clusters can serve as digital biomarkers of aging. Falaschetti et al. [9] further found associations between device-measured activity clusters and chronic conditions, reinforcing the clinical relevance of unsupervised segmentation.

Beyond centroid-based methods like K-Means, probabilistic and functional approaches have been explored. Ensari et al. [10] applied functional mixture models to characterize daily activity trajectories, supporting the use of Gaussian Mixture Models (GMM) for modeling overlapping behavioral groups. A broader review by Farrahi and Rostami [9] emphasized that machine learning techniques—including clustering algorithms—are increasingly central in physical activity, sedentary behavior, and sleep research.

Clustering methods have also been applied in practical and commercial settings. Hartman et al. examined Fitbit use patterns during interventions, showing variability in engagement and activity levels. Additionally, Akansha and So [3] utilized K-Means clustering for Fitbit user segmentation in marketing analytics, demonstrating broader applications of behavioral clustering.

Despite these advances, many studies focus on a single clustering technique. Comparative analysis of multiple algorithms—including K-Means, hierarchical clustering, DBSCAN, and GMM—within the same wearable dataset remains limited. This study addresses this gap by systematically applying and comparing these unsupervised approaches to uncover latent lifestyle and physical activity patterns.

III. Methodology

This study employed a structured analytical pipeline to identify latent lifestyle patterns from wearable sensor data. The methodology included data verification, exploratory analysis, feature selection and transformation, z-score normalization, dimensionality reduction using Principal Component Analysis (PCA), and the application of four unsupervised clustering algorithms (K-Means, Hierarchical Clustering, DBSCAN, and Gaussian Mixture Models). Clustering performance was evaluated using multiple internal validation metrics to ensure robust and interpretable behavioral segmentation. Each process are outlined as follows:

A. Data Collection

The dataset used in this study was obtained from Kaggle, an online data science platform for publicly available datasets. It was uploaded by a user under the name arashnic. The dataset consists of daily physical activity summaries collected from 35 participants wearing Fitbit wearable fitness trackers between March and May 2016. Each record represents one individual's activity behavior for a single day, resulting in 457 daily observations across multiple users. The dataset contains 15 attributes capturing step counts, distance measures, activity intensity durations (very active, fairly active, and lightly active minutes), sedentary behavior, and daily calorie expenditure. Each observation reflects a complete daily behavioral profile. Participants contributed between 8 and 32 days of recorded activity, enabling analysis of intra- and inter-individual variability in lifestyle patterns over time.

B. Data Pre-processing

Data pre-processing was conducted to ensure data quality, reduce redundancy, and enhance the reliability of subsequent clustering analysis. Given the sensitivity of unsupervised learning algorithms to scale, skewness, and correlated features, a structured preparation procedure was implemented to stabilize variance, minimize noise, and optimize feature representation prior to model application.

Initial Sanity Check

The dataset was inspected for missing values, invalid entries, and inconsistent data types. All activity-related numerical features contained complete observations, and no substantial missing data were detected. Identifier and temporal attributes, including participant ID and activity date, were excluded from the feature set as clustering aims to identify behavioral patterns rather than individual identities or temporal trends.

Exploratory Data Analysis

Histograms were used to examine feature distributions prior to transformation, while boxplots were used to assess variability and identify potential outliers. Figure 1 illustrates the raw distributions of daily step counts, distance-based activity measures, intensity-based minutes, sedentary time, and calorie expenditure. Most activity-related variables exhibit strong right-skewness, indicating that the majority of daily observations involve low-to-moderate physical activity, while high-intensity behaviors occur infrequently. This pattern reflects real-world lifestyle heterogeneity where extreme activity levels are rare but influential. Sedentary minutes show wide dispersion, suggesting substantial variability in inactivity across daily observations. These skewed distributions highlight the presence of outliers and non-normality, motivating the application of transformation techniques prior to clustering to prevent extreme values from dominating distance-based algorithms.

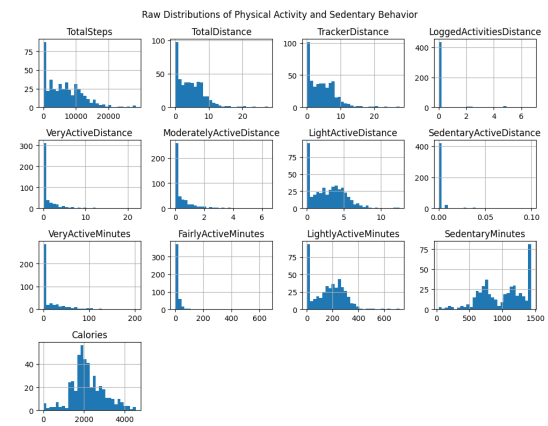

Fig. 1. Raw distributions of physical activity intensity, distance measures, sedentary behavior, and calorie expenditure.

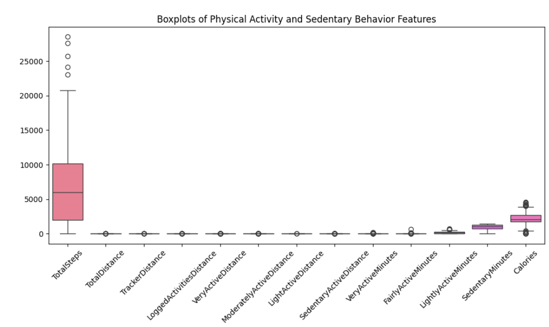

Fig. 2. Boxplots of physical activity and sedentary behavior features highlighting variability and outliers.

Several variables contain long upper tails, confirming the presence of high-activity outliers alongside predominantly low-to-moderate daily movement patterns. Figure 2 further emphasizes the substantial variability across activity-related features and reveals numerous extreme values, particularly in total steps, activity minutes, and calorie expenditure. Sedentary minutes also demonstrate broad interquartile ranges, indicating inconsistent inactivity behaviors across days. These pronounced outliers and unequal feature scales reinforce the necessity of logarithmic transformation and normalization to stabilize variance and ensure balanced feature contribution during clustering.

Strong positive correlations are observed between total steps, total distance, tracker distance, and multiple intensity-based measures, indicating that these variables capture overlapping aspects of movement behavior. Figure 3 presents the correlation structure among activity-related variables. This redundancy suggests multicollinearity within the feature space, which can distort clustering performance and obscure behavioral interpretation. In contrast, sedentary minutes exhibit strong negative associations with movement-related features, reflecting a behavioral trade-off between activity and inactivity. These relationships justified the removal of highly correlated distance-based variables and the retention of behaviorally meaningful activity minutes and sedentary time for subsequent feature selection and dimensionality reduction.

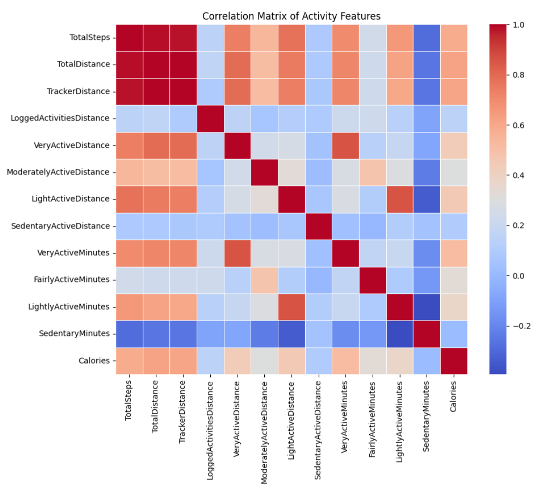

Fig. 3. Correlation matrix of physical activity intensity measures, distance variables, sedentary behavior, and calorie expenditure.

Outlier Detection and transformation

Based on the boxplot analysis in Figure 2, multiple features exhibited extreme upper-tail values, particularly total steps, intensity-based activity minutes, and calorie expenditure. These values reflect rare but legitimate high-activity days that could disproportionately influence distance-based clustering algorithms such as K-Means and hierarchical clustering. Rather than removing these observations, which represent meaningful behavioral extremes, a logarithmic transformation was applied to compress extreme values while preserving relative activity differences. This approach maintains real-world lifestyle variability while reducing the influence of skewed distributions on clustering performance.

Numerical Feature Transformation and Normalization

Given the strong right-skewness, extreme values, and unequal feature scales observed in the raw distributions and boxplots (Figures 1 and 2), a logarithmic transformation was applied to all selected numerical activity variables to reduce distributional asymmetry and stabilize variance. Figure 4 demonstrates that after transformation and normalization, feature distributions became more centered and symmetric, with reduced dominance of extreme high-activity values. Highly skewed variables such as total steps, activity minutes, and calorie expenditure exhibit substantially improved spread and balance compared to their raw forms. Following transformation, z-score normalization was applied to standardize each feature to zero mean and unit variance. Figure 5 confirms that normalized variables now occupy comparable numeric ranges, eliminating scale imbalances between high-magnitude measures (e.g., steps and calories) and lower-magnitude measures (e.g., activity minutes). This combined transformation and normalization process ensures that no single activity dimension disproportionately influences clustering outcomes and that distance-based and probabilistic algorithms operate on a balanced feature space. The resulting normalized dataset provides a stable foundation for dimensionality reduction and unsupervised behavioral segmentation.

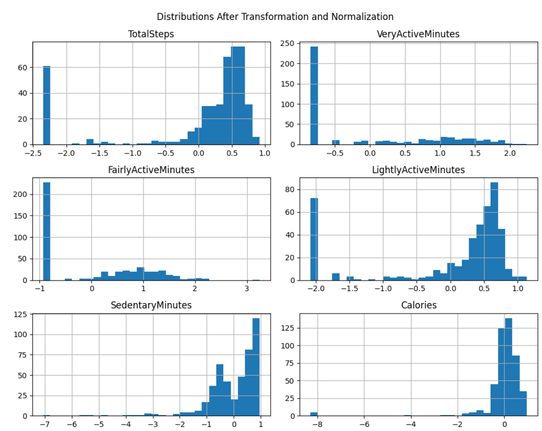

Fig. 4. Distributions of selected activity features after logarithmic transformation and z-score normalization.

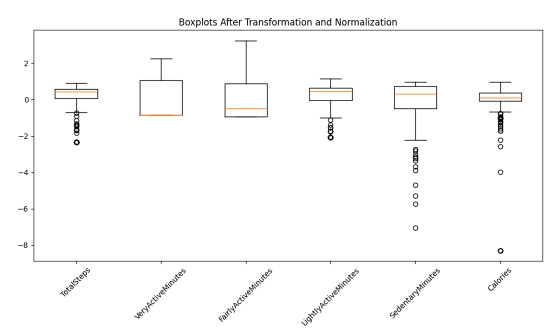

Fig. 5. Boxplots of normalized activity features illustrating reduced skewness and standardized feature scales.

Feature Selection

Correlation analysis (Figure 3) revealed near-perfect multicollinearity among several distance-based variables, including total distance, tracker distance, and step counts, as well as strong overlap between distance and intensity-based activity measures. Retaining all correlated variables would introduce redundancy, reduce interpretability, and potentially distort clustering structure. To address this, the final feature set was restricted to behaviorally meaningful and non-redundant variables: Total steps, Very active minutes, Fairly active minutes, Lightly active minutes, Sedentary minutes, and Daily calorie expenditure

These features collectively represent movement volume, intensity distribution, inactivity duration, and energy output, providing a comprehensive representation of daily lifestyle behavior while minimizing multicollinearity.

Dimensionality Reduction

Principal Component Analysis (PCA) was applied to the normalized feature set to reduce dimensionality while preserving the majority of behavioral variance. PCA transforms correlated activity variables into orthogonal components ranked by explained variance, mitigating multicollinearity and improving computational efficiency. The reduced component space served as the input representation for all clustering algorithms.

C. Experimental Setup

All analyses were conducted using Google Colab with Python-based machine learning libraries. Pandas and NumPy were used for data manipulation, Matplotlib and Seaborn for visualization, and Scikit-learn for preprocessing, PCA, clustering, and evaluation. Clustering was performed on the PCA-transformed feature space to enhance separation and reduce noise. For K-Means and hierarchical clustering, three clusters were explored based on elbow method and silhouette score analysis. DBSCAN parameters were selected through sensitivity testing of neighborhood radius and minimum points. Gaussian Mixture Models were configured with three Gaussian components to ensure comparability across clustering methods.

D. Unsupervised Learning Algorithms Used

To identify latent lifestyle behavior patterns, four unsupervised clustering algorithms were applied: K-Means, Hierarchical Clustering, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), and Gaussian Mixture Models (GMM). These techniques are widely used in physical activity and behavioral research to uncover hidden structures in high-dimensional datasets and wearable-derived metrics [2], [1], [10].

K-Means was selected as the primary algorithm due to its efficiency and suitability for continuous behavioral data. Hierarchical clustering was included to examine nested data structures, DBSCAN to detect density-based clusters and potential outliers, and GMM to model overlapping behavioral distributions through probabilistic assignments. Clustering was performed on the PCA-reduced feature space to enhance separation and computational performance, a strategy commonly adopted in activity pattern analysis [8], [9].

K-Means Clustering

K-Means partitions the dataset into K clusters by minimizing the within-cluster sum of squares (WCSS):

$$J = \sum_{k=1}^{K} \sum_{x_i \in C_k} \|x_i - \mu_k\|^2 \tag{1}$$

Equation 1: K-Means Objective Function

This objective function measures how close each data point is to its assigned centroid; lower values indicate more compact and internally consistent clusters. In behavioral datasets, this helps group individuals with similar activity patterns while maximizing separation from other groups [3], [2].

Hierarchical Clustering

Hierarchical clustering is a technique that organizes data into a nested hierarchy of groups by repeatedly merging or splitting observations based on their similarity. Initially, each data point forms its own cluster, and clusters are then progressively combined according to a selected distance metric and linkage criterion until a complete hierarchical structure emerges. This approach facilitates identifying an appropriate number of clusters by visualizing the distances at which clusters are joined. Unlike other methods, hierarchical clustering does not require specifying the number of clusters beforehand and can uncover the inherent structure within the data. However, it can be computationally demanding and sensitive to the choice of distance measure and linkage method [8], [7]. The following formulas illustrate common linkage methods used to calculate the distance between clusters in hierarchical clustering:

$$d_{12} = \min_{i,j} d(X_i, Y_j) \tag{2}$$

Formula 1. Single Linkage [8]

$$d_{12} = \max_{i,j} d(X_i, Y_j) \tag{3}$$

Formula 2. Complete Linkage [8]

$$d_{12} = \frac{1}{kl} \sum_{i=1}^{k} \sum_{j=1}^{l} d(X_i, Y_j) \tag{4}$$

Formula 3. Average Linkage [8]

$$d_{12} = d(\bar{x}, \bar{y}) \tag{5}$$

Formula 4. Centroid Method [8]

Density-Based Spatial Clustering of Applications with Noise (DBSCAN)

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm that groups observations based on regions of high data density while separating sparse regions as noise. It forms clusters by connecting points that lie within a specified neighborhood radius (epsilon) and satisfy a minimum number of neighboring points. This approach enables DBSCAN to identify clusters of arbitrary shape and effectively detect outliers without requiring the number of clusters to be predefined. However, its performance is sensitive to the selection of epsilon and minimum sample parameters, particularly in datasets with varying densities [1], [2].

Gaussian Mixture Models (GMM)

The Gaussian Mixture Model (GMM) is a probabilistic clustering technique used to identify latent groups within the physical activity dataset. It models the data as a mixture of multiple Gaussian distributions, each representing a distinct cluster. Unlike hard clustering methods, GMM assigns probabilities of membership to each observation, allowing for more flexible cluster boundaries. Model parameters are estimated using the Expectation–Maximization (EM) algorithm, which iteratively updates cluster memberships and distribution parameters until convergence. [10], [9].

The integration of these complementary algorithms improves the robustness of pattern detection by capturing compact clusters, hierarchical relationships, density variations, and probabilistic overlaps. Such methodological diversity has been recommended in recent machine learning studies on physical activity to ensure comprehensive behavioral characterization [2].

E. Training Procedure

Clustering models were fitted on the PCA-reduced dataset. For K-Means, the optimal number of clusters was determined using the elbow method and silhouette analysis. Hierarchical clustering applied Ward's linkage to minimize intra-cluster variance. DBSCAN grouped dense regions while identifying sparse observations. GMM estimated cluster distributions using expectation-maximization.

F. Evaluation Metrics

Clustering performance was assessed using three complementary internal validation metrics: Silhouette Score, Calinski–Harabasz Index, and Davies–Bouldin Index. The Silhouette Score evaluates both cluster cohesion and separation by measuring how similar each point is to its own cluster relative to other clusters, with values closer to 1 indicating well-separated and compact clusters. The Calinski–Harabasz Index quantifies the ratio of between-cluster variance to within-cluster variance, where higher values represent stronger and more distinct clustering structures. The Davies–Bouldin Index measures average cluster similarity, with lower values indicating reduced overlap and improved partitioning quality. Using multiple metrics ensures robust evaluation by capturing different aspects of clustering structure.

G. Comparison of Clustering Algorithms

K-Means Clustering, Hierarchical Clustering, DBSCAN, and Gaussian Mixture Models were systematically compared using the selected validation metrics to determine the most effective segmentation approach for wearable activity data. K-Means consistently achieved the highest Silhouette Scores and Calinski–Harabasz values, indicating superior cluster compactness and separation. Hierarchical clustering produced interpretable groupings but exhibited reduced separation compared to K-Means. DBSCAN demonstrated sensitivity to density variations and identified noise points but resulted in less cohesive clusters. GMM captured overlapping behavioral structures but showed slightly weaker compactness across metrics. Overall, K-Means provided the most stable, interpretable, and well-separated behavioral segmentation while maintaining computational efficiency. Consequently, K-Means was selected as the primary clustering method for lifestyle pattern analysis.

IV. Results and Discussion

This section presents the findings from the clustering analysis conducted on the dataset. Prior to modeling, the features were preprocessed and treated according to the procedures described in the methodology. The results are interpreted to highlight patterns, trends, and insights revealed by the clustering techniques.

A. Dimensionality Reduction

Principal Component Analysis was used to reduce the dimensions while preserving the necessary features.

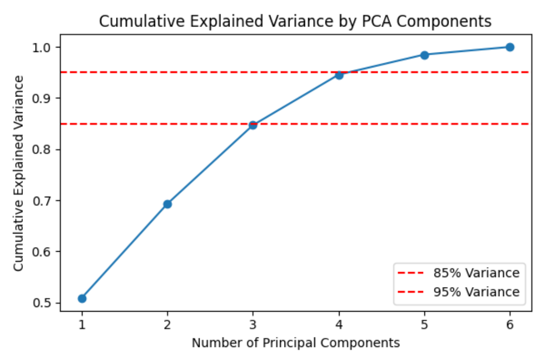

Fig. 6. .Cumulative explained variance across principal components

TABLE I
Explained Variance of Principal Components

Principal Component	Explained Variance %	Cumulative Variance %
PC1	50.88	50.88
PC2	18.43	69.31
PC3	15.39	84.70
PC4	9.88	94.58
PC5	3.92	98.50
PC6	1.50	100.00

Figure 6 presents the cumulative explained variance across principal components. The first principal component (PC1) accounts for 50.88% of the total variance, indicating that overall movement volume and intensity dominate behavioral variation. The second and third components contribute an additional 18.43% and 15.39%, respectively, capturing variations in activity intensity distribution and sedentary behavior. Together, the first three components explain approximately 84.70% of the total variance, demonstrating that most lifestyle behavior information is preserved in a reduced three-dimensional representation. Beyond the third component, additional variance gains become marginal, suggesting diminishing returns from higher dimensions. Therefore, the first three principal components were retained for subsequent clustering to balance information preservation, noise reduction, and computational efficiency.

B. Optimal K Selection

The elbow curve shows a sharp reduction in within-cluster sum of squares (WCSS) from k = 2 to k = 3, followed by a gradual flattening beyond k = 3, indicating diminishing returns in variance reduction with additional clusters. This inflection point suggests that three clusters provide an effective balance between model simplicity and cluster compactness. The silhouette analysis further supports this selection, with the highest silhouette score observed at k = 4, closely followed by k = 3. Although k = 4 yields marginally higher separation, the improvement over k = 3 is minimal while introducing increased model complexity. Considering both structural compactness and interpretability, k = 3 was selected as the optimal number of clusters for subsequent behavioral segmentation. This joint evaluation ensures that the chosen cluster structure captures meaningful lifestyle patterns while avoiding over-fragmentation of behavioral groups.

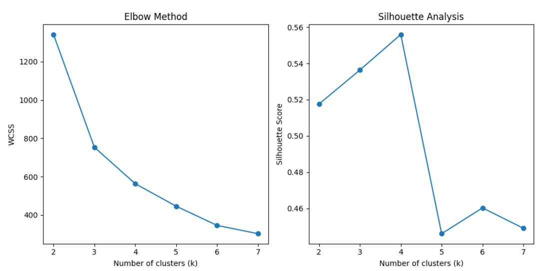

Fig. 7. Elbow Method and silhouette analysis for optimal k value selection

C. Clustering Algorithm Performance Comparison

K-Means achieved the highest Silhouette Score (0.536), indicating the strongest cluster separation and cohesion. Table II summarizes the performance of all clustering algorithms using internal validation metrics. It also produced the highest Calinski–Harabasz Index (472.51), reflecting well-defined and compact behavioral groups. Hierarchical clustering and GMM showed moderate performance but with weaker separation between clusters. DBSCAN performed poorest, with the highest Davies–Bouldin Index (0.943), suggesting substantial cluster overlap and sensitivity to data density variations. Overall, K-Means provided the most stable and interpretable clustering structure and was therefore selected as the primary model for lifestyle behavior segmentation.

D. Cluster Behavioral Profiles and Interpretation

The normalized mean values of selected activity features across the identified clusters.

TABLE II
CLUSTERING PERFORMANCE COMPARISON

Model	Silhouette	Calinski-Harabasz	Davies-Bouldin
K-Means	0.536	472.5	0.711
Hierarchical	0.513	294.2	0.837
DBSCAN	0.526	327.6	0.943
GMM	0.521	432.6	0.730

Cluster 0 exhibits the lowest normalized values for total steps and all intensity-based activity minutes, while showing the highest sedentary time and reduced calorie expenditure.

Cluster 1 displays the highest normalized values across step counts and intensity-based activity measures, accompanied by lower sedentary duration and increased calorie expenditure.

Cluster 2 demonstrates intermediate step accumulation and lightly active minutes, with lower high-intensity activity levels and near-average sedentary behavior. These profiles indicate systematic differences in movement volume, activity intensity distribution, and inactivity duration across clusters.

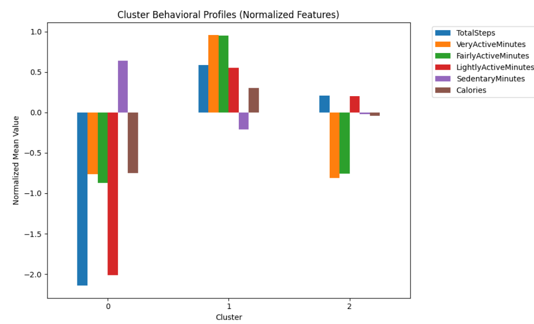

Fig. 8. Normalized behavioral profiles of activity clusters.

E. Behavioral Interpretations and Implications

The identified clusters correspond to distinct lifestyle behavior patterns ranging from sedentary to highly active daily routines.

Cluster 0 reflects predominantly inactive behavior characterized by minimal physical movement and prolonged sedentary periods. This pattern is commonly associated with elevated health risk and low daily energy expenditure.

Cluster 1 represents consistently active behavior with sustained engagement in moderate-to-vigorous physical activity and reduced inactivity time, aligning with movement profiles linked to positive health outcomes.

Cluster 2 captures mixed activity behavior where moderate movement occurs alongside substantial sedentary time, indicating partial adherence to recommended activity levels.

The coexistence of these behavioral segments demonstrates substantial heterogeneity in daily physical activity accumulation among wearable users.

From an applied perspective, such segmentation enables targeted intervention strategies, including inactivity reduction for sedentary groups, structured activity promotion for moderately active individuals, and maintenance-focused programs for highly active populations.

These findings highlight the utility of unsupervised learning for scalable behavioral monitoring and personalized health analytics using wearable sensor data.

V. Conclusion

This study applied unsupervised machine learning techniques to identify latent lifestyle and physical activity patterns from wearable sensor data. Through structured preprocessing, logarithmic transformation, z-score normalization, and dimensionality reduction using Principal Component Analysis, a stable and interpretable feature space was established for clustering analysis. PCA results demonstrated that three principal components retained approximately 84.70% of total behavioral variance, enabling effective dimensionality reduction while preserving meaningful activity information. Based on elbow and silhouette analyses, three clusters were selected as the optimal segmentation structure. Among the evaluated algorithms, K-Means achieved the highest Silhouette Score and Calinski–Harabasz Index, indicating superior compactness and separation of behavioral groups. The resulting clusters represented three distinct lifestyle profiles: predominantly sedentary behavior, moderately active behavior, and highly active behavior. These findings confirm the presence of substantial heterogeneity in daily movement accumulation within wearable datasets. The study demonstrates that unsupervised clustering combined with dimensionality reduction provides an effective and scalable framework for behavioral segmentation using wearable sensor data. Such segmentation can support personalized health monitoring, targeted intervention strategies, and data-driven lifestyle profiling. Future work may incorporate temporal modeling, longitudinal behavior tracking, and larger multi-population datasets to further enhance behavioral pattern discovery and generalizability.

References

[1] F. Pontin, N. Lomax, G. Clarke, and M. A. Morris, "Characterisation of temporal patterns in step count behaviour from smartphone app data: An unsupervised machine learning approach," *International Journal of Environmental Research and Public Health*, vol. 18, no. 21, p. 11476, 2021.

[2] V. Farrahi and M. Rostami, "Machine learning in physical activity, sedentary, and sleep behavior research," *Journal of Activity, Sedentary and Sleep Behaviors*, vol. 3, no. 1, p. 5, 2024.

[3] A. Akansha and S. So, "Revealing sustainable growth for fitbit: A data-driven marketing approach based on k-means clustering and collaborative filtering," in *Proceedings of CSIT, AIRCC*, 2023, pp. 93–108.

[4] World Health Organization, *Guidelines on Physical Activity and Sedentary Behaviour*. Geneva, Switzerland: World Health Organization, 2020.

[5] A. N. Saint-Maurice *et al.*, "Association of step count and intensity with mortality among us adults," *JAMA*, vol. 323, no. 12, pp. 1151–1160, 2020.

[6] D. Ding, B. Nguyen, T. Nau, M. Luo, B. del Pozo Cruz, P. C. Dempsey, Z. Munn, B. J. Jefferis, C. Sherrington, E. A. Calleja, K. H. Chong, R. Davis, M. E. Francois, A. Tiedemann, S. J. H. Biddle, A. Okely, A. Bauman, U. Ekelund, P. Clare, and K. Owen, "Daily steps and health outcomes in adults: A systematic review and dose-response meta-analysis," *The Lancet Public Health*, vol. 10, no. 8, pp. e668–e681, 2025.

[7] S. Nawrin, M. L. McPhee, and J. M. P. Hillsdon, "Examining physical activity clustering using machine learning revealed a diversity of 24-hour step-counting patterns," *Journal of Activity, Sedentary and Sleep Behaviors*, vol. 1, no. 1, pp. 1–14, 2024.

[8] J. Shim, S. Kim, and J. Lee, "Wearable-based accelerometer activity profiles as digital biomarkers of aging using hierarchical clustering," *Scientific Reports*, vol. 13, no. 1, pp. 1–11, 2023.

[9] G. Falaschetti *et al.*, "A cluster analysis of device-measured physical activity behaviours and associations with chronic conditions," *SSM – Population Health*, vol. 15, 2021.

[10] I. Ensari, B. A. Caceres, K. B. Jackman, J. Goldsmith, N. M. Suero-Tejeda, M. L. Odlum, and S. Bakken, "Characterizing daily physical activity patterns with unsupervised learning via functional mixture models," *Journal of Behavioral Medicine*, vol. 48, no. 1, pp. 149–161, 2025.