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 apply-

ing and comparing these unsupervised approaches to uncover latent lifestyle and physical activity patterns.

III. Methodology

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 includes step counts, intensity-based activity durations, sedentary time, and calorie expenditure, capturing both movement and inactivity patterns. Participants contributed between 8 and 32 days of data, allowing analysis of lifestyle variability over time. These objective wearable sensor measurements provide high-resolution behavioral information suitable for unsupervised clustering of physical activity lifestyles.

B. Data Pre-processing

Initial Sanity Check

An initial sanity check was conducted to verify data integrity and consistency. 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

Fig 1 illustrates the distributions of daily step counts, activity intensity minutes, sedentary time, and calorie expenditure. Most activity-related variables exhibit pronounced right-skewness, indicating that the majority of daily observations involve low-to-moderate movement levels, while high-intensity activity occurs infrequently. This imbalance suggests that extreme activity values could disproportionately influence distance-based clustering algorithms. Consequently, transformation techniques were later applied to stabilize variance and prevent outliers from dominating cluster formation.

Sedentary minutes display substantial variability across observations, revealing that some days are dominated by prolonged inactivity while others reflect more balanced movement patterns. This heterogeneity indicates the presence of multiple latent lifestyle behaviors rather than a single dominant activity profile, motivating the use of unsupervised learning to uncover distinct behavioral groups.

Fig 2 presents the correlation structure among activity-related features. Strong positive correlations were observed between total steps and intensity-based activity minutes, indicating that multiple variables capture overlapping aspects of movement behavior. This redundancy suggests that retaining

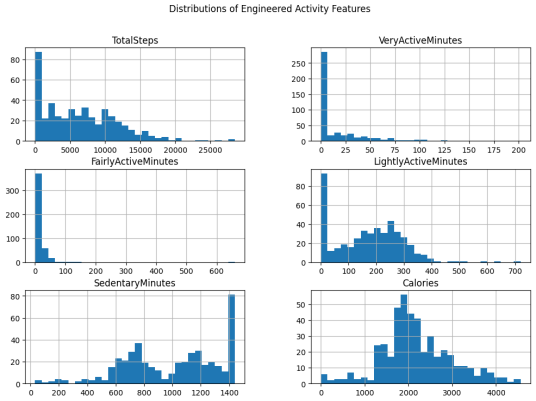

Fig. 1. Distribution of Engineered Activity

all correlated distance-based measures would introduce multicollinearity and obscure cluster interpretability. Therefore, highly correlated distance features were removed in subsequent preprocessing to reduce redundancy and simplify the behavioral feature space.

In contrast, sedentary minutes exhibited strong negative associations with movement-related variables, highlighting an inherent behavioral trade-off between physical activity and inactivity. This relationship justified retaining sedentary time as a core feature to distinguish active and inactive lifestyle profiles.

Together, these observations guided feature selection, transformation, and dimensionality reduction strategies designed to enhance cluster separability and behavioral interpretability.

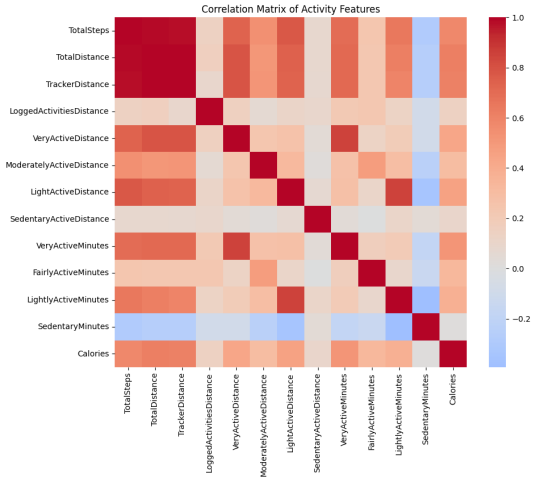

Fig. 2. Correlation matrix of physical activity intensity measures and sedentary behavior.

Figure 3 presents boxplots summarizing the spread, central tendency, and outliers of the selected activity features. Total steps exhibit a wide interquartile range and numerous high-end outliers, indicating substantial variability in daily movement behavior and the presence of occasional extreme activity days. Similar outlier patterns are observed for very active

and fairly active minutes, confirming that intense activity occurs infrequently but reaches high values for a small subset of observations. Sedentary minutes show a comparatively narrower distribution centered at high values, indicating that prolonged inactivity is common across most days. Calorie expenditure displays moderate variability with several extreme values corresponding to highly active days. These distributions reinforce the presence of heterogeneous lifestyle behaviors and justify the application of logarithmic transformation and normalization to reduce the influence of extreme values prior to clustering.

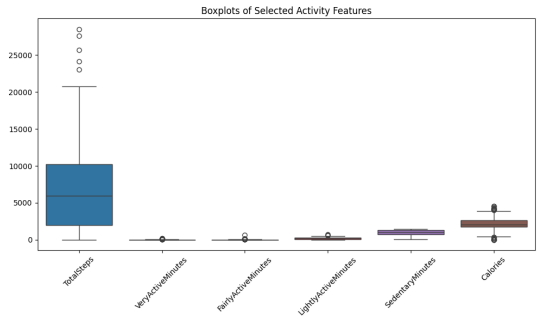

Fig. 3. Boxplots of selected physical activity and sedentary behavior features.

C. Feature Selection and Normalization

Based on correlation analysis, several distance-based variables exhibited near-perfect multicollinearity with step counts and intensity minutes. To reduce redundancy and improve interpretability, only behaviorally meaningful features were retained. The final feature set included total steps, intensity-based activity minutes (very active, fairly active, lightly active), sedentary minutes, and daily calorie expenditure. These variables collectively capture movement volume, activity intensity distribution, inactivity duration, and overall energy output, providing a comprehensive representation of daily lifestyle behavior. Given the right-skewed distributions observed during EDA, a logarithmic transformation was applied to reduce skewness and stabilize variance across activity measures. This transformation improves cluster separability by minimizing the influence of extreme values in distance-based algorithms. All features were standardized using z-score normalization to ensure equal contribution during clustering.

D. Dimensionality Reduction

Principal Component Analysis (PCA) was employed to reduce the dimensionality of the activity feature space while preserving the most significant behavioral variance within the dataset. PCA transforms the original correlated variables into a set of orthogonal components that capture maximal variance in descending order. This approach mitigates redundancy among activity measures, enhances computational efficiency, and improves cluster separability by projecting the data into a lower-dimensional representation.

As shown in Figure 4, the first three principal components captured approximately 85% of the total variance, indicating

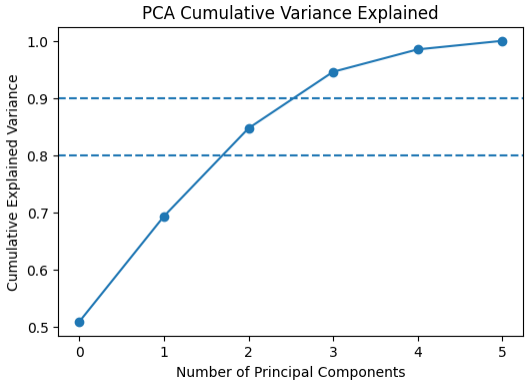

Fig. 4. Cumulative explained variance across principal components.

that the majority of behavioral variability was retained within a lower-dimensional representation. This dimensionality reduction reduces noise and improves clustering efficiency while maintaining interpretability.

E. 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):

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

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 builds a tree-like structure (dendrogram) by iteratively merging clusters based on a linkage criterion. Using Ward's method, the distance between clusters is computed as:

$$D(A,B) = \frac{|A||B|}{|A|+|B|}\|\mu_A - \mu_B\|^2$$

Equation 2: Ward's Linkage Criterion

This formulation minimizes the increase in total within-cluster variance after merging clusters, producing groups that remain as homogeneous as possible. The approach is particularly useful for revealing multilevel behavioral structures, such as subgroups within broader lifestyle categories [8], [7].

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

DBSCAN groups points that are closely packed together while labeling sparse observations as noise. A point is considered a core point if:

$$|N_\epsilon(x_i)| \geq \text{MinPts}$$

Equation 3: DBSCAN Core Point Condition

This condition ensures that clusters are formed only in regions with sufficient density, enabling the algorithm to identify irregularly shaped clusters and detect anomalous behavior patterns. Such capability is valuable in wearable data where outliers may reflect atypical activity or measurement variability [1], [2].

Gaussian Mixture Models (GMM)

GMM assumes that the dataset is generated from a mixture of Gaussian distributions. The probability density function is defined as:

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x \mid \mu_k, \Sigma_k)$$

Equation 4: Gaussian Mixture Model Probability

This equation represents the likelihood that a data point belongs to each cluster, weighted by the mixing coefficients. Unlike hard clustering methods, GMM provides soft probabilistic assignments, allowing individuals to exhibit characteristics of multiple behavioral profiles—a realistic assumption in lifestyle research [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].

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

All four clustering approaches were quantitatively compared across the selected validation metrics to determine the most suitable model for wearable activity segmentation. K-Means consistently achieved the highest Silhouette Score and Calinski–Harabasz Index, indicating superior separation and compactness of behavioral clusters. Hierarchical clustering produced reasonable partitions but exhibited weaker separation, while DBSCAN demonstrated sensitivity to density variations and produced less cohesive clusters. GMM captured overlapping patterns but resulted in slightly lower cluster compactness. Overall, K-Means provided the most stable and interpretable segmentation of lifestyle behaviors, justifying its selection as the primary clustering method.

IV. Results and Discussion

A. Clustering Performance and Visualization

Figure 5 visualizes the clustering structure in PCA space. Three distinct groups are observable. The sedentary cluster appears separated along the negative direction of the primary activity component, while the highly active cluster occupies the opposite region characterized by higher movement intensity. The moderately active group lies between these extremes. The visible separation supports the effectiveness of K-Means in identifying meaningful behavioral segments.

Table I summarizes the distribution of daily activity observations across identified clusters. The moderately active cluster comprised the largest proportion of observations, indicating that most daily behavior patterns involve moderate movement combined with extended sedentary periods. The highly active cluster represented a smaller subset of days, suggesting that sustained high-intensity activity is less prevalent in real-world wearable data. The sedentary cluster accounted for a meaningful proportion of observations, reflecting widespread inactivity across participants. These proportions highlight the heterogeneous nature of daily lifestyle behaviors captured by wearable sensors.

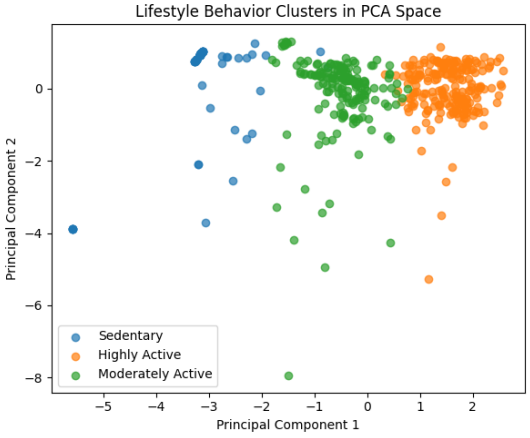

Fig. 5. Lifestyle Behavior Clusters in PCA Space

TABLE I
DISTRIBUTION OF DAILY OBSERVATIONS ACROSS CLUSTERS

Cluster	Number of Observations	Percentage (%)
Cluster 0 (Sedentary)	74	16.19
Cluster 1 (Highly Active)	208	45.51
Cluster 2 (Moderately Active)	175	38.29

B. Cluster Behavioral Characteristics

Figure 6 presents the normalized average values of each activity variable across clusters. Cluster 0 demonstrates extremely low steps and high sedentary time, representing highly inactive individuals. Cluster 1 exhibits high step counts and elevated active minutes, corresponding to physically active participants. Cluster 2 reflects moderate step accumulation and significant light activity but remains characterized by considerable sedentary duration. These behavioral patterns align with known physical activity typologies reported in prior literature [1], [4]. In addition to behavioral averages, cluster distribution analysis revealed that the moderately active cluster comprised the largest proportion of daily observations, suggesting that most individuals exhibit mixed movement patterns characterized by moderate steps but extended sedentary time. The highly active cluster represented a smaller subset of observations, indicating that sustained high-intensity activity is less common in real-world wearable data. The sedentary cluster, while distinct, reflects a meaningful segment of daily behavior patterns associated with prolonged inactivity. These findings align with epidemiological evidence suggesting variability in daily movement accumulation rather than uniform activity patterns [5], [6].

C. Model Evaluation

Figure 7 compares clustering performance across K-Means, Hierarchical Clustering, DBSCAN, and GMM. K-Means achieved the highest Silhouette Score and Calinski-Harabasz Index, indicating superior cluster compactness and separation. Although DBSCAN and GMM demonstrated competitive performance, K-Means provided the most balanced and interpretable segmentation. Therefore, K-Means was selected as the primary clustering algorithm for behavioral analysis.

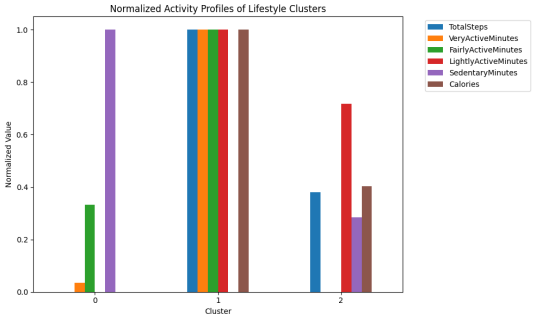

Fig. 6. Normalized Activity Profiles of Lifestyle Clusters

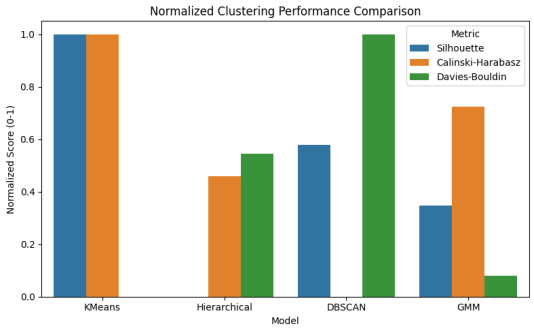

Fig. 7. Normalized Clustering model Performance Comparison

V. CONCLUSION

This study demonstrated the effectiveness of unsupervised machine learning for uncovering latent lifestyle and physical activity patterns from wearable sensor data. Using daily Fitbit activity summaries, a structured pipeline involving exploratory data analysis, feature selection, logarithmic transformation, normalization, and dimensionality reduction was applied to ensure robust and interpretable clustering.

Exploratory analysis revealed strong skewness, substantial behavioral heterogeneity, and high redundancy among movement-related variables, motivating transformation and dimensionality reduction strategies. Principal Component Analysis successfully preserved the majority of behavioral variance within a reduced feature space, enabling improved cluster separability and computational efficiency.

Among the evaluated clustering algorithms, K-Means consistently outperformed hierarchical clustering, DBSCAN, and Gaussian Mixture Models across multiple validation metrics. The resulting clusters revealed three distinct lifestyle profiles: highly sedentary behavior characterized by minimal movement and prolonged inactivity, highly active behavior marked by elevated step counts and intense activity durations, and moderately active behavior reflecting mixed movement patterns with substantial sedentary time. These behavioral typologies align with existing literature on daily movement heterogeneity and physical activity accumulation.

The findings highlight the value of wearable sensor data in capturing real-world lifestyle variability and demonstrate how unsupervised learning can provide actionable behavioral segmentation without predefined labels. Such insights have potential applications in public health monitoring, personalized activity interventions, and digital health analytics.

Despite these strengths, this study is limited by the relatively small sample size and short observation window per participant. Future work may incorporate longer-term longitudinal data, additional behavioral dimensions such as sleep patterns and heart rate, and advanced temporal clustering approaches to capture dynamic lifestyle trajectories. Integrating demographic and health outcome variables may further enhance the interpretability and clinical relevance of identified behavioral groups.

Overall, this research confirms that combining wearable sensor data with robust unsupervised machine learning techniques offers a powerful framework for understanding complex lifestyle behaviors and advancing data-driven health analytics.

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.