Predictive Modeling of Physical Activity Trends of Fitbit Data

Rohit Kumar Bandi Ravikumar, Raghavendra Yadav Golla, and Samah Senbel

Sacred Heart University, Fairfield, CT, USA
{bandiravikumarr, gollar3}@mail.sacredheart.edu,
senbels@sacredheart.edu

Abstract. Wearable technology, particularly devices like Fitbit, offers a wealth of continuous health and fitness data that can be leveraged for predictive analysis. This paper explores the use of machine learning techniques to analyze and predict trends from Fitbit-collected data, focusing on key metrics such as heart rate, physical activity, sleep patterns, and calorie consumption. By applying predictive models to these metrics, we identify patterns in user health behavior and forecast future trends in physical activity and overall well-being. The findings highlight the potential of wearable devices for real-time health monitoring and proactive health management, showcasing how data-driven insights can enhance personalized health outcomes. Our results demonstrate the effectiveness of predictive analytics in wearable health data and suggest pathways for future research in personalized healthcare solutions.

Keywords: Predictive Model, Fitbit data, wearable technology.

1 Introduction

In an era where personal health monitoring has become increasingly prevalent, understanding user activity patterns through fitness tracking devices has emerged as a crucial area of study. The contemporary landscape of health technology is characterized by continuous data collection and monitoring, enabling detailed analysis of individual activity patterns like total steps, active minutes, and distance measurements. This research aims to uncover underlying patterns in user activity data and identify key factors that influence physical activity levels by employing sophisticated data analysis techniques and machine learning algorithms. The goal is to provide insights that could enhance our understanding of user behavior and potentially inform the development of more effective fitness interventions. The study of physical activity patterns through wearable devices has gained significant attention in recent years, with previous research demonstrating the importance of continuous monitoring and analysis of activity data for understanding user behavior and

promoting healthy lifestyle choices. The use of machine learning techniques in analyzing fitness data has shown particular promise in identifying patterns and predicting user behavior, as evidenced by similar studies in the field.

The paper is organized as follows: Sect. 2 summarizes relevant research Fitbit data analysis, Sect. 3 describes our methodology for obtaining, organizing, cleaning, and processing our dataset. Section 4 shows our results of the analysis of Fitbit data, and Sect. 5 concludes the paper.

2 Background

Chelea Deane examines the accuracy and reliability of GPS-tracked activities recorded by the Fitbit Alta tracker and Ionic smartwatch, focusing on their utility in criminal investigations. By analyzing these devices' data, the study identifies potential signs of manipulation or alteration in GPS-tracked activities. The findings aim to assist digital forensic investigators in verifying alibis and constructing timelines in crime scenes, underscoring the importance of wearable device data as forensic evidence.[1]

Zilu Liang assesses the sleep tracking accuracy of the Fitbit Charge 2™, a popular consumer wearable, by comparing it to medical devices. Using a novel validation method with numerical analysis and visual scatter plots for detailed sleep stage comparisons, the study found that the Fitbit Charge 2™ has low accuracy for detecting wakefulness but performs reasonably in identifying light, deep, and REM sleep. The device showed higher accuracy for deep sleep in the first half of the night and REM sleep in the second half. The findings suggest that consumer wearables may not provide high-quality sleep data in natural settings, and future research should explore time-based accuracy variations and segmented modeling to enhance data reliability.[2]

Joseph Williams paper analyzes the Fitbit Versa in forensic contexts, focusing on data recovery and extraction methods useful for criminal investigations. Using Cellebrite UFED and MSAB XRY tools, the study compares logical and physical data extractions from Android 9 and iOS 12 devices. It explores the databases and data types accessible via these methods, assessing the accuracy of Fitbit-recorded data against controlled test scenarios. This work provides a comprehensive overview of data availability, reliability, and potential evidentiary value in forensic investigations involving wearable devices.[3]

Hao Haitao uses the "PARS-3 Physical Activity Grade Measuring Table" and the "Physical Self-Confidence Measuring Table" to assess the physical exercise habits and self-confidence of 327 university students. Results indicate that physical exercise positively impacts both physical health and self-confidence. Exercise frequency significantly influences self-confidence, while endurance and standing long jump strongly affect the relationship between physical health and self-confidence. Additionally, both exercise frequency and duration have a notable impact on physical health.[4]

3 Method

3.1 Data Collection

The data collection process for this research involved aggregating information from multiple fitness tracking devices to create a comprehensive dataset for analysis. The primary focus was on gathering various activity metrics that provide detailed insights into users' physical activities throughout the day. The first key metric collected was total daily steps, which serves as a fundamental indicator of overall physical activity and movement patterns. The distance covered was measured in multiple categories, providing a nuanced view of user mobility - this included total distance traveled, distance covered during very active periods of intense exercise, moderately active periods of sustained movement, and light active periods of casual walking or movement. Timestamp information was carefully recorded to enable temporal analysis and track patterns across different times of day, days of the week, and longer periods. The activity minutes were categorized into four distinct levels of intensity: sedentary minutes representing periods of minimal movement or inactivity, lightly active minutes indicating casual movement like slow walking, fairly active minutes showing moderate physical engagement, and very active minutes representing periods of intense physical activity or exercise. This detailed categorization of activity metrics enables a granular analysis of user behavior patterns and provides a robust foundation for understanding how individuals distribute their physical activity throughout the day. The comprehensive nature of this data collection approach allows researchers to identify patterns, correlations, and trends that could provide valuable insights into user behavior and inform the development of personalized fitness recommendations and interventions.

3.2 Implementation

3.2.1 Data Preparation

To conduct data analysis, the dataset underwent initial preparation where specific columns related to physical activity were selected, including total steps, distance metrics, and various levels of activity (e.g., Very Active, Lightly Active). Standardization was applied using the StandardScaler from scikit-learn to scale the data. This process adjusted all variables to a mean of zero and a standard deviation of one, facilitating comparability across different metrics and ensuring uniformity in distance measurement for clustering.

The initial phase involved selecting relevant columns that represent various aspects of physical activity. These included:

- TotalSteps: Total steps recorded.
- TotalDistance: Total distance covered.
- Distance metrics for varying intensity levels (e.g., VeryActiveDistance, ModeratelyActiveDistance, LightActiveDistance)
- Duration metrics in minutes for each activity level (e.g., SedentaryMinutes, LightlyActiveMinutes, FairlyActiveMinutes, and VeryActiveMinutes).

These columns were selected to capture a holistic view of both the intensity and duration of physical activity. However, differences in the scale of these metrics could introduce bias in clustering and feature importance. For example, TotalDistance (measured in kilometers) and SedentaryMinutes (measured in minutes) exist on different scales, potentially influencing clustering outcomes disproportionately. To mitigate this, **standardization** was applied, which involved centering each variable to a mean of zero and scaling it to a standard deviation of one. This process ensures all features contribute equally to clustering and model training, making comparisons across variables more meaningful.

Using StandardScaler from scikit-learn, each feature was standardized according to:

$$\text{Standardized Value} = \frac{\text{Original Value} - \text{Mean}}{\textit{Standard Deviation}}$$

where the mean and standard deviation were calculated over the training dataset. Standardization is crucial in unsupervised learning and clustering contexts, as it prevents features with larger scales from dominating those with smaller scales.

3.2.2 K-Means Clustering Analysis

Clustering individuals based on physical activity patterns provides insights into different activity levels and behaviors. Here, we employed **K-Means clustering**—a widely used algorithm in unsupervised machine learning, particularly suited for grouping data points based on similarity.

- **Choice of Clusters**: We set n_clusters=3, aiming to distinguish between three levels of physical activity patterns (e.g., low, moderate, high). This choice is based on common classifications in physical activity research, though the number can be adjusted through methods like the Elbow Method to optimize clustering quality.
- **Distance Metric**: K-Means minimizes the **within-cluster sum of squares (WCSS)**, effectively grouping individuals so that those within each cluster are as similar as possible in their physical activity metrics. Each data point was assigned to the nearest cluster centroid, which was iteratively updated to minimize WCSS until convergence.

Using TotalSteps and VeryActiveMinutes as plotting dimensions, a scatter plot was created to visualize the resulting clusters. This visualization elucidates variations in activity patterns within the sample, with each cluster reflecting a different activity profile. For example:

- A cluster with high TotalSteps and VeryActiveMinutes likely represents highly active individuals.
- Conversely, a cluster with low values in both dimensions may indicate less active individuals.

Clustering provides a baseline for further statistical analysis and targeted interventions, where each cluster can be analyzed for specific health outcomes.

3.2.3 Feature Importance Analysis using Random Forest Regression

Feature importance analysis helps identify which activity metrics most influence the target variable, TotalSteps, offering actionable insights for enhancing physical activity. **Random Forest Regression** was chosen due to its ensemble nature, robustness, and ability to handle high-dimensional

data while avoiding overfitting through bootstrapping and aggregation.

- **Target and Predictors**: TotalSteps was selected as the dependent variable, representing the total physical activity outcome. Predictors included distance and duration metrics (TotalDistance, VeryActiveDistance, ModeratelyActiveDistance, etc.), capturing both the quantity and intensity of physical activity.
- **Training and Testing Split**: To evaluate model performance and avoid overfitting, data was split into training and testing sets with an 80/20 ratio, maintaining a random state to ensure reproducibility. Furthermore, to optimize training time, a **50% subset of the training set** was used, allowing efficient modeling without sacrificing predictive accuracy.
- **Model Parameters**: To mitigate overfitting, particularly given the smaller sample subset, the model was configured with n_estimators=10 and max_depth=5. These parameters limit the complexity of individual decision trees within the forest and reduce training time while maintaining accuracy.

During training, the Random Forest algorithm constructed multiple decision trees based on random feature subsets and aggregated the predictions from all trees, a process known as **bagging**. Feature importance was determined by calculating the decrease in node impurity (mean squared error) each feature contributed across all trees. Features that led to greater impurity reductions were assigned higher importance scores.

- **Feature Importance Interpretation**: After training, feature importance values indicated the relative influence of each activity metric on TotalSteps. Metrics with higher importance scores directly inform which activity types (e.g., very active or lightly active minutes) significantly impact the total step count. This insight can guide tailored interventions—for instance, encouraging increased very active minutes if they are found to correlate strongly with higher step counts.

3.2.4 Model Deployment

he trained Random Forest model was serialized using **joblib** into a pickle file (random_forest_model.pkl). Serialization allows for the preservation of model structure, parameters, and learned weights, making it possible to reload and apply the model on new datasets without retraining. In practice, this allows the model to be deployed in a production environment or real-time analysis setup, where new data can be fed through

the model to predict TotalSteps or analyze feature importance dynamically.

This pipeline—spanning data preprocessing, clustering, feature analysis, and deployment—provides a robust framework for analyzing physical activity data, uncovering insights that can inform healthcare policies, fitness recommendations, and personalized interventions. The combined approach of clustering and feature importance analysis supports not only data-driven profiling but also prescriptive insights into specific activity patterns that contribute most significantly to total physical activity levels.

4 Results

The figure-1 depicts the daily step trends of five users over a one-month period, illustrating variations in physical activity levels across individuals. Each user is represented by a different color, with the x-axis showing the dates and the y-axis indicating total daily steps. One prominent observation is the variability in step counts between users. For instance, User 3, represented in green, maintains a relatively stable step count, consistently achieving around 10,000 to 15,000 steps daily. In contrast, other users, such as User 2, marked in orange, display significant fluctuations, including a notable peak on May 1st, where steps exceed 30,000—a value much higher than typical daily counts. This spike suggests a specific event, perhaps an intense exercise day or a lengthy walk, that led to an unusually high step count.

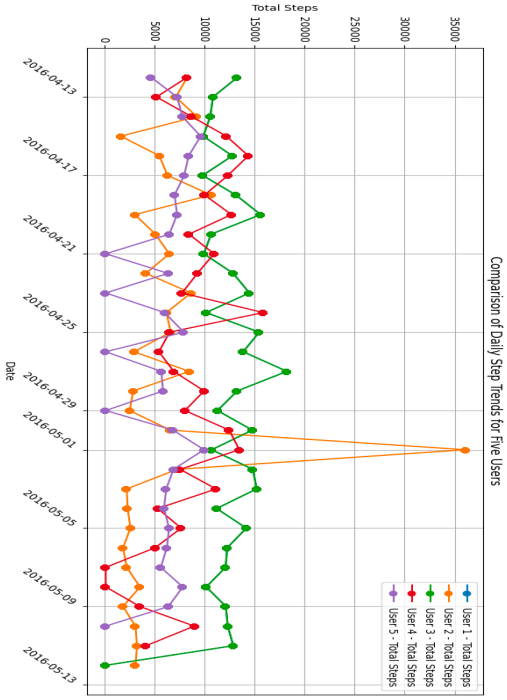

Figure 1 – Comparison Of Daily Step Trends for Five Users

Furthermore, the figure-1 reveals distinct overall activity levels among users. User 3 (green) generally has the highest step counts, reflecting a more active daily routine, whereas Users 5 (purple) and 1 (blue) exhibit lower activity levels, often recording fewer than 5,000 steps per day. This could indicate less engagement in physical activities or a lifestyle with fewer opportunities for movement. Despite fluctuations in activity, there is no clear seasonal or cyclical trend, as the data shows irregular peaks and dips rather than a consistent pattern of high or low activity days. This analysis highlights individual differences in physical activity habits, which could be valuable for personalized fitness recommendations or for understanding broader trends in daily movement across diverse lifestyles.

The Figure-2 provides an insightful analysis of the relationships between various physical activity metrics. Each cell reflects the correlation coefficient between two variables, with values ranging from -1, indicating a strong negative correlation, to 1, indicating a strong positive correlation. The color gradient enhances readability, with dark red representing strong positive correlations and dark blue indicating strong negative correlations.

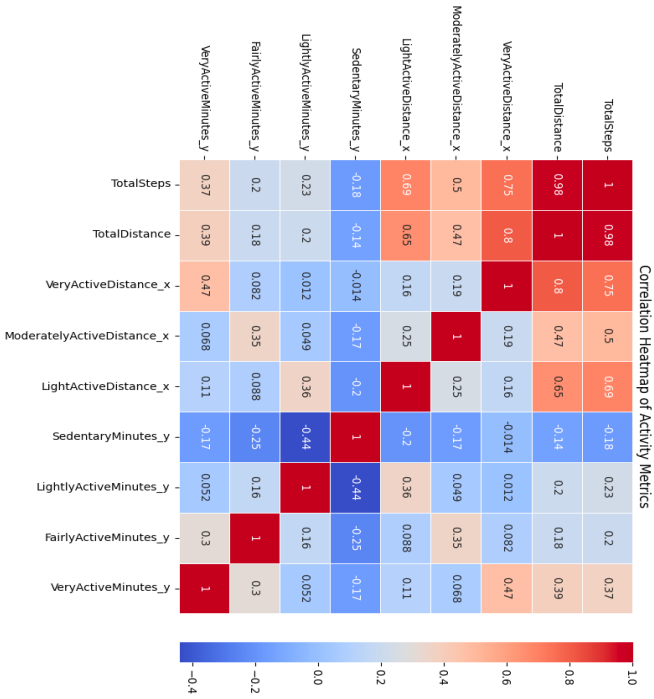

Figure 2 - Correlational Heatmap of Activity Metrics

One of the most notable findings is the strong positive correlation between **TotalSteps** and **TotalDistance** (0.98), showing that as individuals take more steps, the total distance they cover also increases significantly. This close association is expected, as higher step counts directly contribute to greater distances. Additionally, **VeryActiveDistance** correlates strongly with both **TotalDistance** (0.8) and **TotalSteps** (0.75), suggesting that periods of vigorous activity are a major contributor to overall step counts and distance covered. This pattern implies that engaging in very active minutes significantly boosts daily totals in both distance and steps.

Moderate correlations are also observed, particularly between **LightActiveDistance** and **TotalSteps** (0.69), indicating that light-intensity activities, such as casual walking, meaningfully impact overall step counts. Furthermore, **LightlyActiveMinutes** correlates moderately with **LightActiveDistance** (0.36), reflecting that more time spent on light activities translates into a higher distance covered during these activities. This moderate correlation underscores the role of lower-intensity activities in contributing to overall physical movement, even if their impact is less than that of vigorous activities.

On the other hand, **SedentaryMinutes** shows negative correlations with most physical activity metrics, particularly with **LightlyActiveMinutes** (-0.44), indicating that as sedentary time increases, light activity time decreases. Negative correlations with **TotalSteps** and **TotalDistance** further highlight that increased sedentary behavior is associated with reduced physical activity levels. These findings suggest that while both high- and low-intensity activities positively contribute to total steps and distance, sedentary behavior detracts from these totals, reinforcing the importance of minimizing inactivity for overall health. This correlation analysis offers a clear perspective on how different activity levels interact and impact overall physical activity, providing a basis for strategies aimed at increasing physical engagement by emphasizing specific activity types.

The Figure-3 displays the 7-day rolling average of total steps for five users over a month, smoothing daily fluctuations to highlight longer-term trends in physical activity. Each user is represented by a different color, and the y-axis shows the rolling average of total steps.

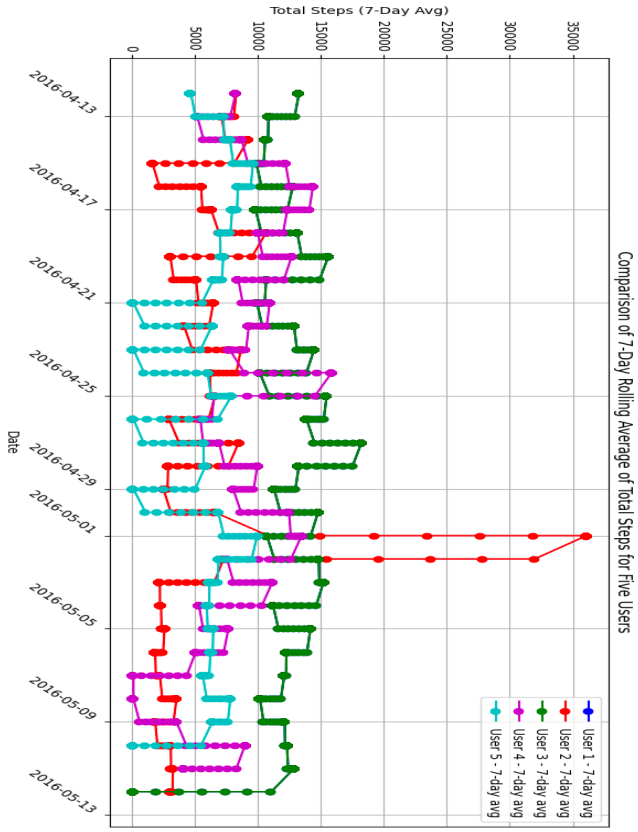

Figure 3 - Comparison of 7-Day Rolling Average of Total Steps for Five Users

One noticeable pattern is the relatively stable activity levels for some users, such as User 3 (green), whose rolling average consistently hovers around 10,000-15,000 steps, indicating a steady engagement in physical activity. In contrast, User 2 (red) experiences a significant spike in their 7-day average around May 1st, where the rolling average climbs above 30,000 steps. This spike likely reflects a single day of exceptionally high activity, which elevated their average for the surrounding week.

Other users, such as Users 1 (blue), 4 (purple), and 5 (cyan), generally maintain lower rolling averages, with steps often staying below 10,000. This suggests a lower level of adaily physical activity for these individuals. The rolling average data for these users shows slight fluctuations but remains relatively low, which may indicate more consistent but limited activity levels.

The Figure-4 shows clustering based on activity metrics, specifically using **Total Steps** on the x-axis and **Very Active**

Minutes on the y-axis. Each point represents an individual data point, color-coded by cluster, with three clusters identified in blue, red, and gray.

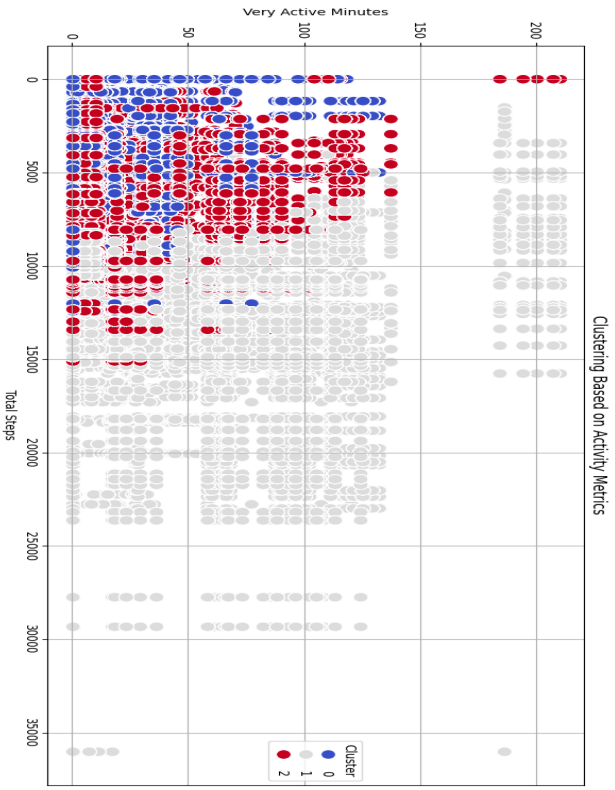

Figure 4 - Clustering Based on Activity Metrics

The clusters reveal distinct activity patterns among the groups:

1. **Cluster 0 (Blue)**: This cluster is concentrated in the lower-left section of the plot, with low to moderate values for both Total Steps and Very Active Minutes. Individuals in this cluster generally engage in lower-intensity activities, often recording fewer than 10,000 steps and limited very active minutes. This may represent a group with sedentary or light activity levels.
2. **Cluster 1 (Gray)**: This cluster is the most widespread and spans across activity profiles, including some outliers with very high values in either Total Steps or Very Active Minutes. The broad distribution indicates diversity within this cluster, with members that may occasionally engage in high-intensity or high-duration activities but do not consistently achieve high levels of either metric.
3. **Cluster 2 (Red)**: Concentrated in the lower-middle range of Very Active Minutes and moderate to high Total Steps,

this cluster suggests individuals who may achieve higher step counts through sustained low-to-moderate activity rather than intense bursts. These participants generally accumulate steps but do not spend extended periods in very active minutes.

The Figure 5 & Figure 6 is a **Total Steps Predictor** app built with Streamlit, which allows users to predict their total step count based on input activity metrics. The app provides sliders for inputting values related to various physical activity metrics, such as Total Distance, Very Active Distance, Light Active Distance, Sedentary Minutes, Lightly Active Minutes, Fairly Active Minutes, and Very Active Minutes. Each slider has a range tailored to typical values for each activity metric, allowing users to set precise activity levels based on their expected or observed data.

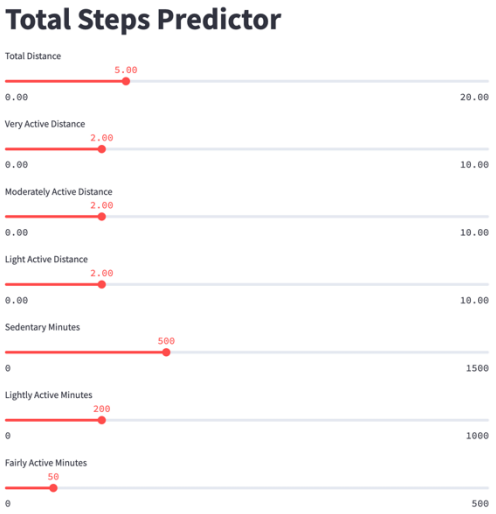

Figure 5 – Interface 1

Figure 6 – Interface 2

The app uses a pre-trained Random Forest model to make predictions based on these inputs. As the user adjusts the sliders, they define the activity profile, which is then processed by the model to output a predicted total step count, displayed at the bottom of the interface (e.g., "Predicted Total Steps: 7316"). This feature offers users an interactive and intuitive way to understand how different activity levels impact their daily steps, providing actionable insights for managing or achieving fitness goals.

5 Conclusion

In this paper, we presented a comprehensive analysis of physical activity data, using clustering, feature importance, and predictive modeling to explore patterns in daily step counts and activity levels. By segmenting users into distinct clusters based on activity metrics, we revealed unique behavioral profiles that highlight the diversity of physical engagement across individuals. The clustering analysis identified groups ranging from low to high activity, offering valuable insights for personalized fitness recommendations tailored to different lifestyle patterns.

The feature importance analysis, powered by a Random Forest model, underscored the impact of various activity types on total step counts. Specifically, very active minutes emerged as the most influential factor, suggesting that intense physical activity plays a crucial role in achieving higher daily step totals. However, moderate and light activities also contribute significantly, emphasizing the value of incorporating a mix of activity intensities for overall physical health. Meanwhile, sedentary time negatively correlated with physical activity, reinforcing the importance of reducing inactivity to maintain or improve fitness levels.

Finally, the predictive model developed through this analysis enables individuals to estimate their total step count based on specific activity inputs. The interactive Streamlit app allows users to experiment with different activity settings and understand how changes in their behavior might affect their overall activity levels. This tool provides practical, data-driven insights, empowering users to make informed decisions about their physical activity habits.

References

1. Atheer Almogbil; Abdullah Alghofaili; Chelsea Deane; Timothy Leschke; Atheer Almogbil; Abdullah Alghofaili: The Accuracy of GPS-Enabled Fitbit Activities as Evidence: A Digital Forensics Study. In proceedings of the 2020 7th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2020 6th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom).
2. Zilu Liang; Mario Alberto Chapa Martell: Combining Numerical and Visual Approaches in Validating Sleep Data Quality of Consumer Wearable Wristbands. In proceedings of the 2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops).
3. Joseph Williams; Áine MacDermott; Kellyann Stamp; Farkhund Iqbal: Forensic Analysis of Fitbit Versa: Android vs iOS. In Proceedings of 2021 IEEE Security and Privacy Workshops (SPW).
4. Hao Haitao : Research on the Relationship between Physical Exercise, Physical Health and Physical Self-confidence of University Students. In Proceedings of 2011 International Conference on Future Computer Science and Education.
5. Jong-Ho Yoon, Seung-Hyun Lee, Kyoungchul Kong: Walking Pattern Classification and Walking Distance Estimation Algorithms Using Gait Phase Information. In IEEE Transactions on Biomedical Engineering, 2012.
6. Yong-Jae Kim, Seung-Ho Hyon, Atsuo Takanishi: Measurement of Human Walking and Generation of Humanoid Walking Pattern Based on Human Walking. In IEEE/RSJ International Conference on Intelligent Robots and Systems, 2008.
7. Yong-Jae Kim, Seung-Ho Hyon, Atsuo Takanishi: Online Walking Pattern Generation for Humanoid Robot with Compliant Motion Control. In IEEE/RSJ International Conference on Intelligent Robots and Systems, 2019.
8. Majid Khadiv, Alexander Herzog, S. Ali A. Moosavian, Ludovic Righetti: Walking Control Based on Step Timing Adaptation. In IEEE Transactions on Robotics, 2020.
9. Shuuji Kajita, Fumio Kanehiro, Kenji Kaneko, Kazuhito Fujiwara, Kazuhiko Yokoi, Hirohisa Hirukawa: High Frequency Walking Pattern Generation Based on Preview Control of Zero-Moment Point. In IEEE International Conference on Robotics and Automation, 2003.
10. Zhanpeng Jin, Edward J. R. Petraglia, Behnam Ayub, Michael E. Kiani, Maria Luisa Nascimento: "A Wearable System for Gait Training in Stroke Rehabilitation." In IEEE Transactions on Neural Systems and Rehabilitation Engineering, 2017.
11. Daniel Roggen, Alberto Calatroni, Mauro Rossi, Thomas Holleczek, Gerhard Troster: "Collecting Complex Activity Datasets in Highly Rich Networked Sensor Environments." In IEEE International Conference on Networked Sensing Systems, 2012.