Predicting Sleep Duration and Sleep Quality Using Machine Learning and Fitbit Wearable Data:

An applied analysis of wearable activity data for sleep modeling

Grace Okoro

2025

Introduction

Sleep is extremely important to human health. Quality sleep can determine cardiovascular function, metabolic regulation, cognitive performance, and psychological well-being. Chronic insufficient sleep has been linked to elevated risks of hypertension, diabetes, obesity, impaired attention, and long-term mortality (Medic et al., 2017). Because of these outcomes, accurate and scalable sleep measurement is critical in both clinical practice and population health.

Most traditional sleep analysis relies on polysomnography (PSG) as the standard for evaluating sleep stages and identifying sleep disorders. However, this PSG approach requires overnight observation in specialized laboratories, access to trained technicians, and is extremely costly ($2,000 per night or even more). These barriers severely limit accessibility and make long-term sleep monitoring impractical for most individuals. In recent years, wearable devices such as Fitbit fitness trackers have expanded opportunities for continuous, non-invasive, daily sleep monitoring at low cost. Although these wearables can provide insight into sleep behavior, the proprietary algorithms used to classify sleep states are not transparent and can vary in accuracy (de Zambotti et al., 2019). In this case, machine learning can be helpful by uncovering meaningful behavioral patterns in wearable data and clarifying how daily activity relates to sleep.

This project originally aimed to classify sleep stages using the DREAMT dataset, a multimodal wearable dataset collected alongside laboratory PSG recordings (PhysioNet, 2025). DREAMT provided a massive amount of data for 100 participants and seemed to be well suited for deep-learning approaches such as convolutional neural networks and long short-term memory networks. However, I found that the computational demands of training deep-learning architectures on DREAMT exceeded available system resources. To maintain feasibility while preserving the scientific goals of the project, I pivoted to a public Fitbit dataset with daily activity measures and nightly sleep duration (Arashnic, 2020). This allowed me to develop an ML pipeline for sleep prediction while creating a foundation for future work with complex multimodal datasets.

The purposes of this study are threefold: 1. develop and evaluate ML models that predict sleep duration and sleep quality from FitBit's activity features, 2. determine which behavioral patterns are most strongly linked to sleep outcomes, and 3. create a scalable framework that can later be applied to signal-level datasets such as DREAMT

Problem Description and Data

This study used the publicly available Fitbit Fitness Tracker Dataset, which includes about one month of data for 30 adults (Arashnic, 2020). I found three files to be relevant to this project: *dailyActivity_merged.csv* (daily steps, distances, calories, and activity intensities), *minuteSleep_merged.csv* (minute-level labels of asleep, restless, or awake), and *weightLogInfo_merged.csv* (BMI and weight, which was not used in the analysis).

Because the sleep data was recorded at the minute level, I started with aggregating those values into daily summaries. For each day, I calculated the total number of minutes labeled as "asleep", along with the total number of minutes spent in bed. The difference between these two values determined the minutes spent awake during the sleep period. These daily sleep summaries were merged with daily activity data using participant ID and date, producing substantial daily records with both activity features and sleep outcomes.

The Fitbit dataset provided a variety of activity-based predictors. These include distance measures (TotalDistance, TrackerDistance, LoggedActivitiesDistance), intensity-based distances

(VeryActiveDistance, ModeratelyActiveDistance, LightActiveDistance, SedentaryActiveDistance), time-based activity metrics (VeryActiveMinutes, FairlyActiveMinutes, LightlyActiveMinutes, SedentaryMinutes), and total Calories. These variables were used as predictors for both regression and classification tasks.

I then defined two target variables. For regression, the outcome was TotalMinutesAsleep, which is a continuous measure of nightly sleep duration. For classification, I defined sleep quality by comparing each night's sleep duration to the sample median of 420 minutes (7 hours). Nights at or above the median were labeled "good sleep," and nights below the median were labeled "poor sleep."

Methods for Data Analytics and Machine Learning

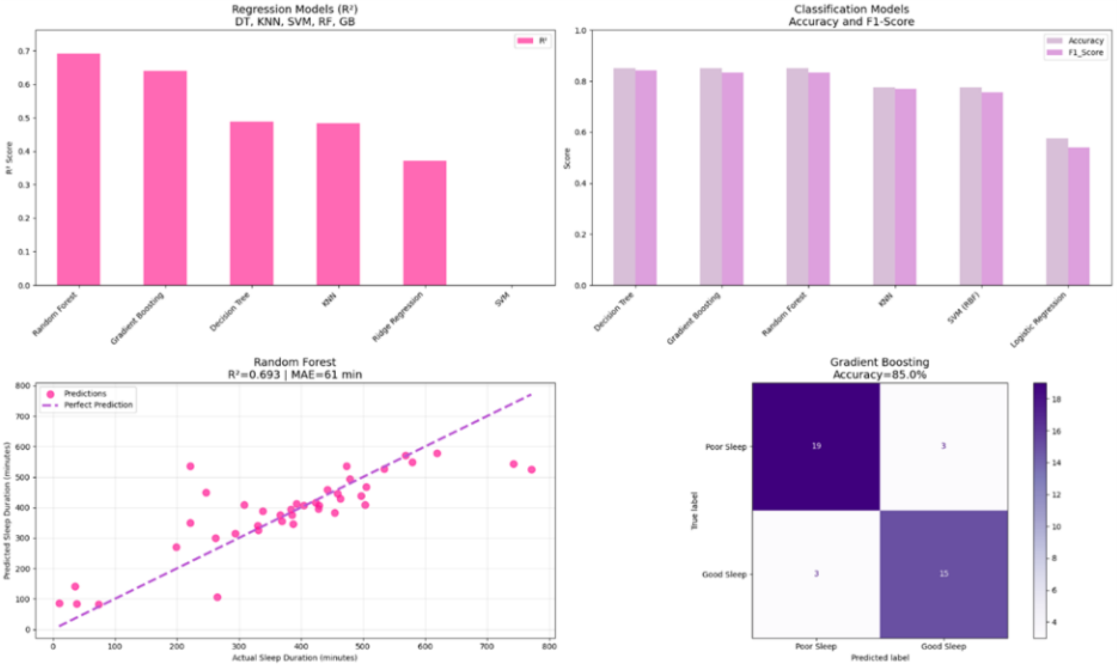
Before I could do any model training, I had to preprocess the data, which included converting timestamps to date format, aggregating minute-level sleep metrics into daily values, and merging sleep and activity data by participant and date. Days that had no sleep or activity data were removed. The final dataset was split into training and test sets using an 80/20 split with a fixed random seed (42) for reproducibility, and all predictors were standardized using z-scores to avoid bias in model training.

The final feature set consisted of the 12 activity variables, with missing values imputed as zeros when no activity was logged for any particular metric. Six models were trained for each task. I started with fewer but wanted to cover more of the machine learning models discussed this semester. For classification I used Decision Tree, K-Nearest Neighbors (KNN), Support Vector Machine (RBF kernel), Logistic Regression (liblinear solver), Random Forest, and Gradient Boosting. For regression I used Decision Tree Regressor, KNN Regressor, Support Vector

Regressor (RBF kernel), Ridge Regression, Random Forest Regressor, and Gradient Boosting Regressor.

In order to determine the best performing classifier, I used a weighted scoring system that combined test accuracy, F1-score, and cross-validation (CV) stability, with weights of 0.35 for accuracy, 0.35 for F1, and 0.30 for CV stability. This prioritized models that performed well on the test data while maintaining consistency across cross-validation, reducing the risk of overfitting. For regression models, performance was evaluated using R^2, which measures the proportion of variance explained by the model, and MAE, which quantified the average prediction error in minutes.

Figure 1. Model Comparisons



Evaluation Methods

The classification models were evaluated using test accuracy, precision, recall, F1-score, five-fold CV accuracy, and the gap between test and CV accuracy as an overfitting indicator. A confusion matrix also examined how well good versus poor sleep nights were identified.

Regression models were assessed with R^2 and MAE and visualized using predicted-versus-actual scatterplots. Additionally, Pearson correlations and p-values examined the relationship between each activity variable and sleep duration, while feature importance scores from ML models demonstrated which behaviors mattered most.

Results

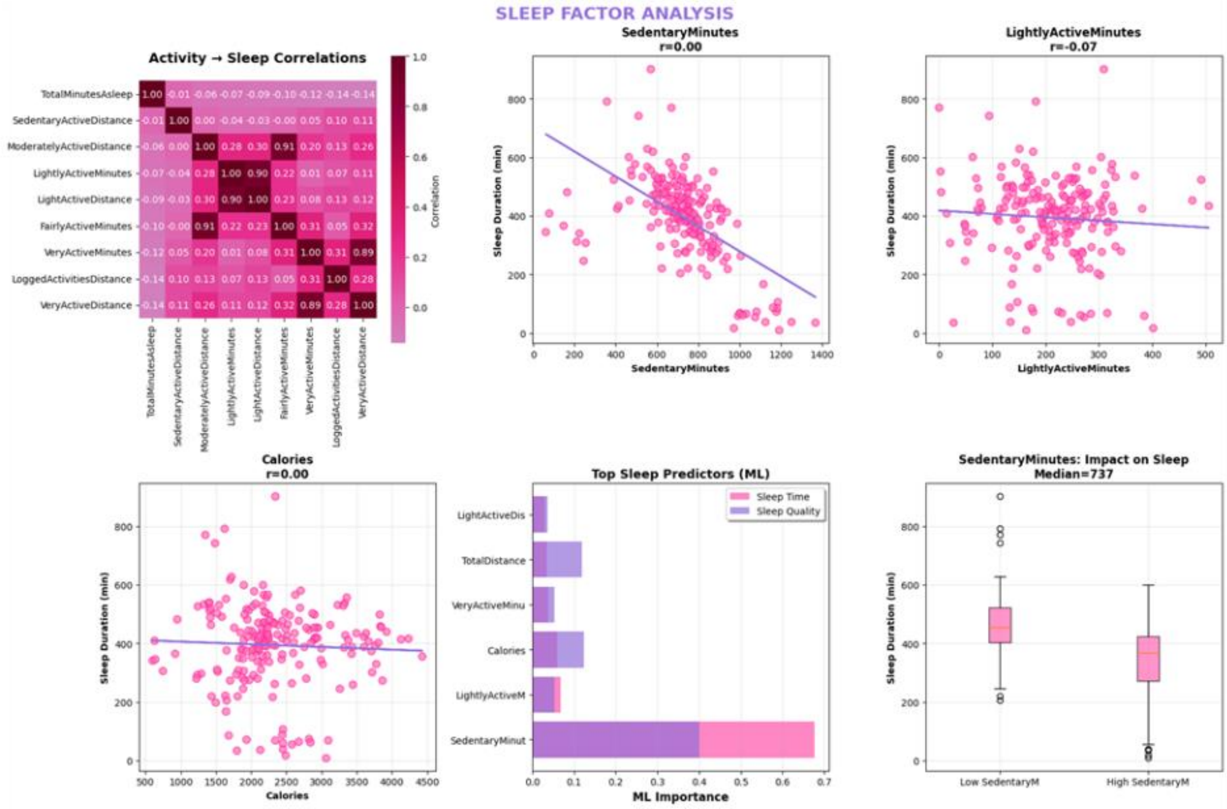
Among the classification models, Gradient Boosting performed best, achieving 85.0% accuracy, an F1-score of 0.833, and cross-validation stability of 81.8% (having the highest combined weighted score 0.834). The confusion matrix indicated that the model correctly identified 86% of poor-sleep nights and 83% of good-sleep nights, which suggests that there are potential real-world applications such as behavioral coaching or personalized sleep alerts.

For regression, the Random Forest was the top model, explaining 69.3% of the variance in sleep duration (R^2=0.693) and achieving a mean absolute error of about 60.9 minutes.
This indicates that Fitbit activity data can account for a substantial portion of variability in sleep duration (from day to day).

SedentaryMinutes emerged as the dominant behavioral predictor in both tasks. In the regression model, it contributed 67.7% of total feature importance, and in the classification model, 40.1%. All other predictors had single digit importance values. SedentaryMinutes showed a strong negative correlation with sleep duration ($r = –0.582$, $p < .001$), meaning that more sedentary time was associated with shorter sleep. TotalDistance had a weaker but significant association (r = –0.150, p = .035), while other activity metrics were not significantly related to sleep duration.

These relationships have some practical behavior implications. Participants with higher sedentary time slept about 78 minutes less per night than those with lower sedentary time. Reducing sedentary time by just 30 minutes per day corresponded to roughly 18 additional minutes of sleep, while a reduction of 60 sedentary minutes was linked to more than half an hour of sleep, which is three to four more hours of sleep per week. These relationships demonstrate how simple daily behavior change can influence sleep health.

Figure 2. Sleep Factor Analysis



Discussion

This study shows that daily activity data from consumer wearables can be used to predict sleep duration and sleep quality with reasonably high accuracy. The Gradient Boosting classifier achieved 85% accuracy for distinguishing good versus poor sleep nights, which is on par with or better than many wearable-based models described in the literature (de Zambotti et al., 2019; Lee et al., 2025; Wang et al., 2025).

SedentaryMinutes prevailed as the strongest behavioral predictor, which suggests that overall sedentary time may be more relevant for sleep outcomes than time spent in vigorous or moderate physical activity. While intense exercise may benefit sleep in general, this dataset suggests that simply reducing sedentary periods and increasing general movement may have a larger impact on sleep duration and quality.

The Random Forest regression model explained nearly two-thirds of the variance in sleep duration, exceeding performance reported in other studies of consumer wearables (de Zambotti et al., 2019). This indicates that relatively simple daily summary metrics contain actionable insights for better sleep.

Because DREAMT integrates multimodal wearable streams aligned with PSG, this same machine learning framework, scaled and extended with deep learning, could support tasks such as sleep stage classification or detection of irregular sleep patterns. This would allow for more precise modeling of sleep architecture while preserving the practical advantages of wearable monitoring (Lee et al., 2025).

Limitations

The sample size was relatively small (n = 30), which reduces the generalizability of the findings. Additionally, the dataset did not include demographic or behavioral covariates such as age, caffeine consumption, stress levels, screen time, or bedtime consistency, which likely influence sleep and could improve model performance. Fitbit's measurements are less accurate than PSG, which means sleep duration estimates may contain measurement errors. The analysis also relied on daily summary metrics rather than minute level time series, which restricts the ability to capture behavioral patterns throughout the day and how their timing affects sleep. Lastly, some activity metrics, such as logged distances, may underestimate true activity levels due to irregular tracking.

Future Work

Future work would build on these findings. Applying the pipeline to the DREAMT dataset would enable the use of multimodal sensor inputs and support deep learning models such as CNN–LSTM hybrids for sleep stage classification and prediction of sleep disturbances (PhysioNet, 2025; Wang et al., 2025). Incorporating additional behavioral and contextual variables such as stress, caffeine intake, bedtime regularity, and screen exposure would likely improve predictive accuracy. Integrating measures of sleep regularity and circadian rhythm would also provide a more complete picture of sleep health and may improve early detection of health risks.

Conclusion

This study shows that ML models can use Fitbit activity data to predict nightly sleep duration and classify sleep quality with practical accuracy. SedentaryMinutes was the most influential behavioral predictor, highlighting that even mild physical activity can improve sleep. The ML pipeline developed here not only supports short-term health recommendations but also provides a foundation for future work with multimodal datasets such as DREAMT, where more advanced models could further improve personalized sleep guidance, risk detection, and digital interventions.

References

Arashnic. (2020). Fitbit fitness tracker data [Data set]. Kaggle. https://www.kaggle.com/datasets/arashnic/fitbit

de Zambotti, M., Cellini, N., Goldstone, A., Colrain, I., & Baker, F. (2019). Wearable sleep technology in clinical and research settings. Sleep Health, 5(6), 687–694.

Lee, H., et al. (2025). Predicting sleep quality with digital biomarkers and artificial neural networks. Frontiers in Psychiatry. https://www.frontiersin.org/articles/10.3389/fpsyt.2025.1591448

Medic, G., Wille, M., & Hemels, M. E. (2017). Short- and long-term health consequences of sleep disruption. Nature and Science of Sleep, 9, 151–161.

PhysioNet. (2025). Dataset for real-time sleep stage estimation using multisensor wearable technology. https://physionet.org/content/dreamt/

Wang, W., et al. (2025). Algorithmic development towards field-based sleep monitoring. DukeSpace. https://dukespace.lib.duke.edu/items/230c2e2f-7621-481c-943a-b1bcf98f4236