

Tracking Depression with Wearable Device Data

Darshika Verma

Harrisburg University of Science and Technology

Abstract

This study looks at how to detect depression using physical activity data from wearable actigraphy devices. Using an openly available dataset, the analysis starts with summary statistics and behavioral feature extraction: average activity, standard deviation, maximum activity, and total activity per day. These features help differentiate between healthy and depressed participants. I use machine learning models such as Logistic Regression, Random Forest, and XGBoost with reproducible participant-level splits (60:20:20 train/validation/test ratio). I evaluate the models through F1 scores, ROC, and AUC curves. Hyperparameter tuning with Optuna and GPU acceleration significantly improves performance. Random Forest achieves the best macro F1-score of 0.74 and AUC of 0.82. I use SHAP analysis to improve interpretability, which confirms that activity patterns are relevant behavioral indicators. The results show that ensemble and boosting methods, along with strong experimental controls, effectively differentiate depression. This supports the use of sensor-based analytics in mental health research.

Keywords: wearable devices, depression detection, activity patterns, machine learning, SHAP analysis

Tracking Depression with Wearable Device Data

Depression is a common and disabling mental health disorder that affects individuals, families, and societies worldwide. The common features of all the depressive disorders are sadness, emptiness, or irritable mood, accompanied by somatic and cognitive changes that significantly affect the individual's capacity to function (Chand & Arif, 2023). The World Health Organization (2024) estimates that over 264 million people around the world suffer from depression, making it a major cause of disability. The impact of depression goes beyond the individual, affecting families, communities, and economies through lost productivity, higher healthcare costs, and a decline in quality of life (Bains & Abdijadid, 2023). Reports indicate that 1 in 8 people globally has a mental illness (World Health Organization, 2024), and within a year of the COVID-19 pandemic, more than 26% of individuals reported experiencing symptoms of depression or anxiety. Depression greatly contributes to the global burden of disease and is linked to various other conditions, including anxiety, substance abuse, heart disease, and diabetes (Harmer et al., 2024). It can seriously affect a person's ability to function socially and at work, and in severe cases, may lead to suicide.

Identifying individuals with depression symptoms is vital for delivering the right interventions and treatments. Early detection of depression and related mental health issues is important because untreated conditions can become life-threatening (Uban et al., 2021). Researchers use a range of methods and tools for this purpose, from self-report questionnaires to digital technologies that analyze speech patterns and social media activity. Kour (Uban et al., 2021) notes that distinguishing between depressed and non-depressed individuals is challenging due to the lack of practical methods. Moreover, there are not enough resources and trained healthcare professionals to treat depression. Machine learning (ML) and deep learning (DL) are among the methods used for diagnosing depression. Researchers have long been interested in identifying depression from user-generated content online, providing mental health professionals with better screening tools (Bucur et al., 2023). Recent studies

show that AI tools aim to change mental healthcare by delivering remote assessments of depression risk using behavioral data collected by sensors in smartphones. While these tools can accurately predict heightened depression symptoms in small, similar groups, they have shown less accuracy in larger, more diverse populations (Adler et al., 2024).

Diagnosis also heavily relies on self-report tools. These include clinical interviews or questionnaires like the PHQ-9, which are common in both clinical and non-clinical environments. While these tools can be useful, they are subjective. Their effectiveness depends on how a person responds and their ability to convey their mental state. Consequently, individuals who cannot or choose not to share their feelings may go unnoticed. Although tools like the PHQ-9 work well in terms of measurement, many self-administered screening tools face usability problems. They are often too long, complex, or dependent on literacy, which makes them difficult to use effectively in primary care. There is also no agreement on the best tool to use, and even the most effective tools do not ensure better detection unless there are proper care systems in place (Miller et al., 2020).

In the last twenty years, there has been growing interest in wearable technology. Wearables include any devices that people can wear, such as pedometers, headphones, and virtual reality headsets. Smartwatches are the most common type of wearable, accounting for about 30% of the total market (Statista, 2023). Prominent examples include the FitBit and Apple Watch. Smartwatches collect a wide range of health data about users, like heart rate and physical activity. They provide easy access to this information through their displays or companion apps. Interest in monitoring personal fitness and mental health is on the rise (Cruz et al., 2024). However, mobile mood-tracking apps are often underused, and users may avoid logging negative feelings (Schueller et al., 2021). Wearables can continuously gather extensive data and have previously been used to identify the emotions users experience (Saganowski et al., 2020; Shu et al., 2020). Additionally, wearables can collect this data discreetly, making them suitable for long-term tracking with minimal disruption. The role of wearables in

mental health care has been a subject of interest for some time. While existing reviews focus on different wearable sensors and features that can inform predictive models (Abd-Alrazaq et al., 2023; Ahmed et al., 2023; Kang & Chai, 2022; Lee et al., 2021), this review concentrates more on the movement of patients.

Literature Review

Recent advancements in digital health show the promise of wearable technologies for monitoring depression symptoms by passively collecting physical activity data. Watanabe and Tsutsumi (2022) created a deep learning model that uses physical activity data from workers. This model achieved a classification accuracy of 76.3% in predicting levels of psychological distress. Their findings show that working conditions, the timing of activities, and routine patterns can significantly improve prediction models for depression and anxiety. Similarly, Price et al. (2023) conducted a year-long study that used wearable movement and sleep data to track symptom variability in individuals with major depressive disorder. Their research indicates that depression is not constant; fluctuations in symptoms over time are vital for identifying relapse risk and customizing interventions. Together, these studies emphasize how passive, long-term behavioral data, like activity intensity, sleep cycles, and work context, can provide valuable insights into mental health monitoring and personalized care strategies.

The study of vocal patterns has become a powerful and affordable way to identify depressive symptoms in both clinical and non-clinical environments. Taşçı (2024) suggested a multilevel hybrid feature extraction model that analyzes speech audio using statistical and wavelet-based techniques to detect depression. Using the MODMA dataset, the model achieved an accuracy of 94.63%, showing it is highly effective in diagnosis while still being computationally efficient. By identifying subtle phonetic markers like speech hesitancy, monotone delivery, and low vocal energy, the system captures important behavior signs linked to major depressive disorder. In addition, Huang et al. (2024) introduced a deep learning framework that uses the Wav2Vec 2.0 pre-training model to automatically extract features

related to depression from raw speech. Their method detected changes in vocal pitch, rhythm, and emotional tone, highlighting the diagnostic importance of acoustic data. Together, these findings strengthen the potential of speech-based AI tools in improving early depression screening through passive and scalable methods.

Recent research has shown the effectiveness of wearable and sensor-based technologies for tracking mood and anxiety disorders. Lee et al. (2025) examined the diagnostic and monitoring abilities of GPS-derived mobility features using Fourier transform analysis in individuals with bipolar disorder (BP) and major depressive disorder (MDD). Their study highlighted several key GPS indicators, including location variance (LV), transition time (TT), and entropy, which varied significantly between mood states. They found that depressed states were associated with lower LV (OR 0.975, 95% CI 0.957-0.993; P=.008), TT (OR 0.048, 95% CI 0.012-0.200; P<.001), and entropy (OR 0.662, 95% CI 0.520-0.842; P=.001).

This emphasizes how daily mobility patterns can serve as noninvasive markers of mood. Similarly, Dao et al. (2025) used wearable electrooculography (EOG) and electrodermal activity (EDA) signals to find physiological indicators of state anxiety. Their feature-based model achieved an accuracy of 98.17% and showed that specific EOG and EDA metrics, like opening phase energy, signal height, Hjorth activity, and spectral entropy, played a major role in predicting anxiety states. These findings illustrate how multimodal wearable data can enable real-time, context-aware detection of mood and anxiety disorders.

With the growing use of wearable technology in mental health research, recent studies have focused on how these tools can help detect and monitor depression in daily life. Shui et al. (2025) collected daily physiological data using wristband wearables from individuals diagnosed with depression. Their model, which used heart rate, skin conductance, and body acceleration, achieved up to 90% accuracy in identifying depression with just six hours of data. This suggests that subtle body signals can help detect mental health symptoms in everyday life. Similarly, Abd-Alrazaq et al. (2023) conducted a

systematic review and meta-analysis of 54 studies and found that wearable-based artificial intelligence models performed well in detecting depression, with average accuracy as high as 89%. However, they also noted that performance varied depending on the algorithms and devices used, showing that more research is still needed to improve consistency. Supporting these findings, Borghare et al. (2024) discussed how wearable technology helps monitor depressive symptoms over time, increases patient engagement, and enables personalized treatment. They emphasized how wearables are now being used in remote care, giving doctors real-time insights and helping patients stay connected to their mental health goals.

Depression symptoms change over time. Single assessments may not fully capture these variations. Price et al. (2023) looked at this by analyzing one year of wearable movement and sleep data along with PHQ-9 scores from 939 participants. Their study found that using both domain-driven and complete feature inclusion models could moderately predict long-term symptom changes ($r = 0.33$ and $r = 0.39$). This confirms that depression does not follow a fixed pattern and must be tracked over time. Similarly, Rykov et al. (2024) studied older adults with mild cognitive impairment. They showed that physiological data collected from wearable sensors, such as heart rate variability and skin conductance, could predict depression severity scores with high accuracy. Their deep learning models achieved correlations of $r = 0.73$ for depression severity, $r = 0.67$ for mood disorder symptoms, and $r = 0.69$ for mild behavioral issues. This supports the value of combining physiological signals with deep-learning features to monitor mental health. However, Sun et al. (2023) pointed out critical challenges in using smartphone and wearable features, such as mobility, sleep, and phone use, to predict depression severity.

Their study found that some features worked better for short-term analysis, like sleep onset time, while others, such as wakefulness after sleep onset, were more useful for tracking over time. They also emphasized the importance of participant engagement and the need to understand how behavior

patterns shift between high and low PHQ-8 periods. These studies show that wearable and mobile health data can help detect depression and shed light on how it evolves. While wearable AI technologies have great potential for screening and monitoring anxiety and depression, some key limitations exist. Abd-alrazaq et al. (2023) conducted a scoping review that examined 69 studies involving AI-enabled wearable devices for mental health.

They found that most studies concentrated on diagnosing depression using data from physical activity, sleep, and heart rate, primarily through wrist-worn devices like the Actiwatch. However, none of the studies looked at using wearable AI for treatment purposes. The review also pointed out gaps in previous research methods. Earlier reviews often left out the AI aspect, did not conduct thorough database searches, and targeted limited age groups or specific disorders. For instance, many prior studies did not include databases like PsycINFO or IEEE Xplore. Some studies focused only on children or adults, which led to biased conclusions. Moreover, a common limitation across studies was the narrow focus on biosignals. Many studies depended only on electrocardiogram or electroencephalogram data without considering multimodal data streams. The authors stressed the urgent need for stronger, more inclusive, and AI-integrated research frameworks to increase the reliability, generalizability, and clinical relevance of wearable AI in mental health.

In clinical settings, wearable technology is becoming an important tool for improving traditional psychiatric care. Fedor et al. (2023) show how long-term data from wearables can lead to deeper conversations between patients and clinicians by providing objective insights that may be missed in subjective self-reports. For example, sleep data collected by a wearable device helped reveal disturbed sleep patterns that the patient did not initially discuss in therapy. The study notes that wearables do not replace clinical evaluations; instead, they support them by confirming patient- reported symptoms with real-time physiological data. The authors argue that including this passive data collection in routine

clinical care can help with early detection, ongoing monitoring, and more personalized treatment strategies for people with depression.

While existing research shows that wearable technologies can detect depression by examining physical activity, many studies look at broad behavioral features, such as sleep duration, GPS location changes, or heart rate variability. These features can provide valuable insights, but they often depend on complex multimodal datasets or deep learning models, which can be hard to understand or replicate in real-world clinical environments. Additionally, some models rely heavily on proprietary datasets, making it tough to validate or apply their findings to different populations. Another significant limitation in previous research is the lack of attention to daily movement patterns derived from raw actigraphy data. Most studies either aggregate data over long periods or focus on specific time frames, such as sleep at night. This approach misses out on day-to-day behavior changes that might indicate shifts in mental health. Moreover, there are few studies that quantify these patterns using basic statistical indicators, like total activity or standard deviation, which could make the results more understandable and useful for clinicians, especially in low-resource settings where technical tools or labeled multimodal data may not be available.

Moreover, despite the progress in wearable AI for mental health, there are still significant challenges. Many studies depend on deep learning models (Taşçı, 2024; Huang et al., 2024). While these models can be accurate, they often lack clarity and are difficult to use in real-world clinical settings. Several models rely on multimodal or proprietary datasets (Abd-Alrazaq et al., 2023; Rykov et al., 2024), which makes it hard to reproduce results across different groups. Additionally, most existing research targets average behaviors (like total sleep time and average heart rate) or data gathered at night (Sun et al., 2023). This approach misses daytime physical activity patterns, which could serve as clearer indicators of depression.

To address the limitations of subjective self-reports and the complexities of current AI-based screening models, this study seeks to create simple, understandable behavioral indicators from daily physical activity patterns recorded by wrist-mounted actigraphy devices. This research looks at an interpretable and objective way to detect signs of depression using physical activity data collected from wrist-worn Actiwatch devices. The dataset came from the Simula Research Laboratory (Simula Research Laboratory, 2021) and includes high-resolution movement records from participants diagnosed with either unipolar or bipolar depression, along with healthy control subjects. Each participant wore a wrist-mounted Actiwatch, which recorded movements at a frequency of 32 times per second.

These readings were averaged and saved as a single movement value for each minute, known as activity_counts_per_min, reflecting the individual's physical activity level at that moment. To simplify analysis and improve clarity, the data was combined at the daily level rather than using minute-by-minute measurements. The dataset included participant identifiers, date stamps, and a binary label indicating whether a participant was in the depressed group (1) or the healthy group (0). To create meaningful behavioral indicators from the raw movement data, four statistical features were developed: average activity, standard deviation of activity, maximum activity, and total activity per day. Average activity measures the mean daily movement and indicates general energy levels.

The standard deviation shows how varied a participant's movements were during the day, suggesting behavioral consistency or inconsistency. Maximum activity reflects the most active minute in a day, while total activity aggregates all minute-by-minute activity scores for that day. These metrics were calculated using Python's data processing libraries, where daily data was grouped by participant and summarized with standard statistical functions. To determine if there were significant differences in activity patterns between the depressed and healthy groups, I performed an independent two-sample t-test for each of the four features.

The results revealed clear patterns, with healthy individuals generally showing higher average and total activity levels, as well as greater movement variability. In contrast, the depressed group displayed lower overall activity, more consistent movement patterns, and less variability. These findings align with clinical symptoms of depression, such as psychomotor slowing, fatigue, and decreased engagement in daily activities. A key strength of this research is its focus on interpretability. While machine learning has the potential to help spot depression through social media, most current studies face issues like bias, overfitting, poor language processing, and lack of transparency (Cao et al., 2024). There is considerable opportunity for improvement to make these models reliable and useful in real-world mental health care. In contrast, this study uses a clear analytical framework based on easy-to-compute and explainable metrics.

By keeping the model simple, the analysis becomes accessible to a wider audience, including clinicians, developers, and researchers needing actionable and clear results. Furthermore, this research highlights the promise of using real-world, passively collected wearable data as a supplementary tool for early detection and ongoing monitoring of mental health issues. Unlike traditional screening tools that heavily depend on self-reporting or clinician-administered interviews, wearable devices enable constant, non-intrusive monitoring. As these devices become more integrated into daily life, they provide a valuable chance to track behavioral signs of depression over time, even before individuals feel ready or able to express their symptoms. This approach addresses a critical gap in mental healthcare by offering a non-verbal, objective way to assess mental health.

In summary, the findings indicate that daily movement patterns captured through wearable actigraphy can reveal significant differences between depressed and non-depressed individuals. Using straightforward statistical summaries—along with easy-to-understand visualizations—provides a scalable, transparent, and effective way to support early detection, monitoring, and personalized interventions in mental health care.

Method

Description of the Data

The dataset used in this research comes from the Simula Research Lab. It was created for a study on behavioral patterns in individuals diagnosed with depression. This dataset is the foundation of my project titled “Tracking Depression with Wearable Device Data”. The goal is to find out if physical activity patterns recorded by wearable devices can effectively differentiate between individuals with depression and those who are mentally healthy.

Table 1

Dataset Variables and Descriptions Used in Depression-Related Activity Analysis

Fields	Interpretation
timestamp_minute	The exact minute when the activity was recorded (e.g., 2004-09-28 09:30:00)
date	The date on which the movement was tracked (e.g., 2004-09-28)
activity_counts_per_min	The amount of movement in 1 minute (unitless count; higher = more movement)
participant_id	The ID of the person (e.g., condition_1, control_5)
label	Depression label: 1 = depressed, 0 = healthy
recorded_days	How many days the person wore the activity tracker
gender_code	Gender of the person: 1 = male, 2 = female
age_group	Age of the person, shown as a range (e.g., 45-49)
affective_disorder_type	Type of depression: 1 = bipolar I, 2 = bipolar II, 3 = unipolar
melancholia_status	If the person shows signs of melancholia (a deep sadness type of depression)
inpatient_status	Hospitalization status: 2 = inpatient (hospital), 1 = outpatient (not hospitalized)
education_level	Level of education (e.g., 6-10 = 6 to 10 years of schooling)
marital_status	Marital status: 1 = married, 2 = not married
employment_status	Work status: 1 = employed, 2 = unemployed
madrs_score_before	Depression score (MADRS) at the beginning of tracking
madrs_score_after	Depression score (MADRS) at the end of tracking

Note. Variables names are as coded in the original dataset. Higher values indicate greater movement or activity, except where otherwise specified. All categories values are explained in interpretation column.

The study includes two different groups of participants. The first group has 23 individuals diagnosed with unipolar or bipolar depression, while the second group consists of 32 individuals without any diagnosed mental health issues, serving as the control group. All participants wore a wrist device called an Actiwatch. This medical-grade wearable can record even the smallest movements at a frequency of 32 Hz, capturing data 32 times per second. These micro-movements were collected every

minute to create a value called `activity_counts_per_min`. Higher values indicate more movement, while lower or zero values show inactivity or minimal physical engagement. Each participant wore the Actiwatch for five to twenty-two days, with an average of about fourteen days of continuous tracking. The dataset therefore provides a rich temporal resolution, showing a time-stamped and detailed view of daily movement patterns across individuals with different mental health profiles.

The dataset includes three main files. The file `condition_all.csv` has raw movement data for participants with depression, while `control_all.csv` contains movement data for mentally healthy individuals. The third file, `scores.csv`, provides demographic and clinical metadata, such as gender, age group, education level, marital and employment status, type of depressive disorder (if applicable), hospitalization status (inpatient or outpatient), presence of melancholic features, and depression severity scores measured with the Montgomery–Asberg Depression Rating Scale (MADRS) at the start and end of the observation period. These three files are combined using participant identifiers to create a single dataset that merges high-resolution activity data with relevant clinical information. This integration allows for a thorough analysis of how physical movement patterns might reflect underlying mental health conditions.

Description of Variables to Be Studied

To address the research goal of comparing physical activity between depressed and healthy individuals, set of variables are created and selected in two main categories: behavioral features from raw movement data and contextual variables from participant metadata.

The behavioral features summarize daily physical activity levels. Four main metrics are calculated which are the independent variables. The first is average activity (`avg_activity`), which shows the mean activity counts per minute for a participant on a specific day, reflecting their overall movement level. The second is standard deviation (`std_activity`), which captures the variability or consistency of activity throughout the day. Lower values suggest steady activity, while higher values

indicate more variation. The third is maximum activity ('max_activity'), which indicates the highest single-minute activity count recorded in a day, providing insight into peak physical exertion. The fourth variable, total activity ('total_activity'), is the sum of all minute-by-minute activity counts over a 24-hour period, offering a complete measure of daily movement. These engineered features are easy to compute and interpret. They are also supported by existing research as meaningful indicators of behavioral changes linked to depression.

In addition to behavioral features, several contextual variables are included from the 'scores.csv' metadata. These consist of a binary label for mental health status, where 1 represents depressed individuals and 0 represents healthy controls. Demographic details are added like gender (coded as 1 for male and 2 for female), age group (in categorical ranges), and educational background (based on years of formal education). Clinical variables include the type of affective disorder (Bipolar I, Bipolar II, or Unipolar Depression), melancholia status (present or absent), inpatient versus outpatient classification, and socioeconomic data like marital and employment status. MADRS scores is included from the start and end of the observation period to assess the severity and progression of depressive symptoms. Including contextual variables like age, gender, and MADRS scores helps explain differences in physical activity behavior. Age and gender influence baseline movement levels. MADRS scores confirm connections to depression severity. These variables improve the analysis by linking behavioral patterns to clinical insights.

Data cleaning, preprocessing, and analysis

The data cleaning and processing used Python along with libraries such as Pandas, NumPy, Matplotlib, and Seaborn.

Step 1: Data Integration

The first step involved reading all three datasets, condition_all.csv, control_all.csv, and scores.csv, into separate data frames. Each dataset was checked for structural consistency. Then, labels

were added to identify participants diagnosed with depression (label = 1) and those in the control group (label = 0). After labeling, the condition and control datasets were combined to create one movement data table. This movement dataset was then merged with the demographic and clinical metadata from scores.csv using the unique participant identifier (participant_id). The result was a dataset that connected high-frequency movement patterns with individual clinical and demographic profiles.

Step 2: Data Cleaning

A series of cleaning operations were carried out to prepare the data for analysis. Redundant columns, like the variable number that duplicated the participant_id, were removed to cut down on noise. Categorical variables that were originally coded numerically, such as gender as 1 or 2, were converted into descriptive labels, like 'Male' and 'Female', to make them easier to understand. Data types were carefully checked and corrected. For example, timestamp variables were changed from strings to datetime objects to allow for temporal grouping. Missing values were evaluated across all columns, and suitable strategies, such as imputation or removing rows, were applied based on the context and effect of the missing data. Additionally, column names were standardized to improve clarity; one example is renaming the generic activity field to the more descriptive activity_counts_per_min.

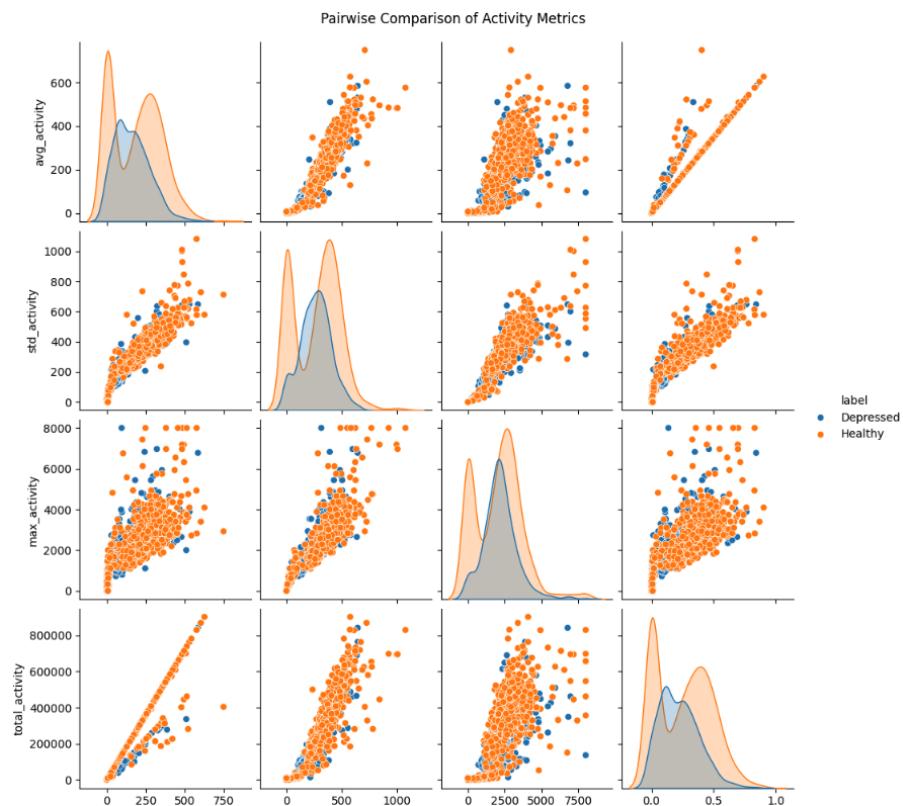
Step 3: Feature Engineering

To change the high-resolution minute-by-minute movement data into a format suitable for behavioral analysis, data is gathered daily for each participant. This involved organizing the data by participant_id and date, then calculating key summary statistics for the activity_counts_per_min variable. The new features created included the daily average activity (avg_activity), standard deviation (std_activity), maximum activity value (max_activity), and total activity counts per day (total_activity). This process produced a new table where each row showed a participant's behavioral activity profile for a specific day. It made later analysis much simpler while keeping essential behavioral signals intact.

These four metrics were explored through pairwise comparisons, as shown in the first uploaded image. This pairplot illustrates how depressed and healthy participants are distributed across various combinations of the four activity metrics. Notably, the density curves along the diagonal show that healthy individuals (orange) generally have higher activity levels, especially in avg_activity and total_activity, compared to depressed individuals (blue).

Figure 1

Pairwise Comparison of Activity Metrics



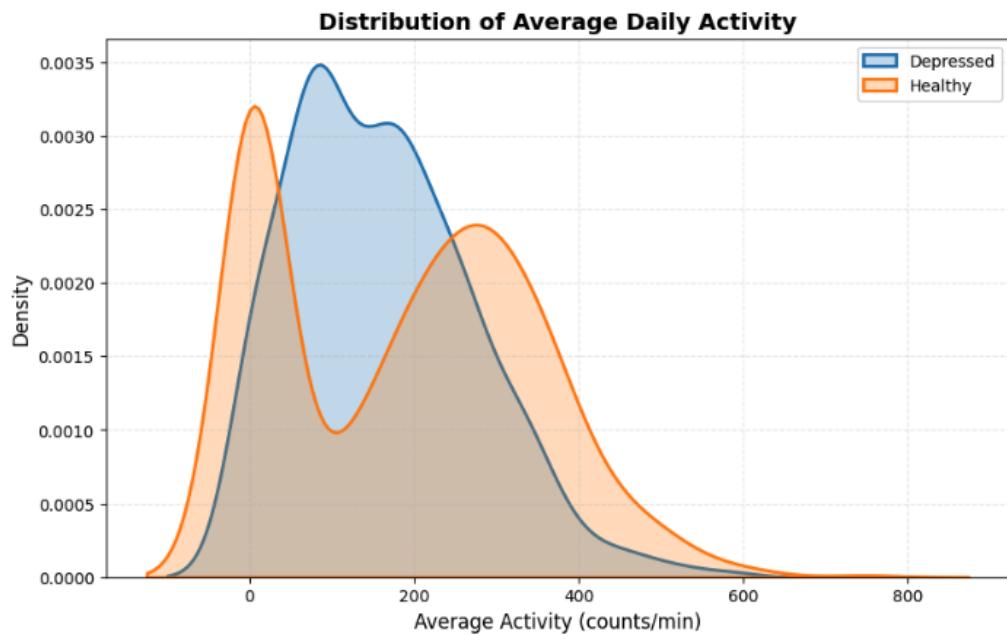
Note. Each point represents an individual's summary activity metric. Activity metrics were computed from Simula Research Laboratory (2021).

Step 4: Exploratory Data Analysis

To explore the behavioral differences in physical activity between depressed and healthy individuals, I used several data visualizations with Seaborn and Matplotlib. A boxplot comparing average daily activity (avg_activity) showed that depressed participants generally had lower median activity levels and less variability. In contrast, healthy individuals exhibited higher activity, a greater range, and more outliers. To investigate the distribution of activity levels, I created a kernel density estimate (KDE) plot.

Figure 2

Distribution of Average Daily Activity Levels for Depressed and Healthy Participants



Note. Lower and less variable activity levels are observed in the depressed group compared to the healthy group. Data source: Simula Research Laboratory (2021).

On the other hand, the healthy group had a wider, bimodal distribution with higher peaks in activity. This visual difference supports the idea that reduced and less variable activity relates to depression.

Step 4: Statistical Analysis – T-Test

To determine if there is a significant difference in daily physical activity between individuals diagnosed with depression and those in a healthy control group, an independent two-sample t-test was conducted using the `avg_activity` variable. This variable reflects average movement per minute throughout the day. Given the potential for unequal variances, Welch's t-test was used.

Figure 3

T-test Comparison of Average Daily Activity Levels between Depressed and Healthy Groups



Note. Average daily activity was significantly lower in the depressed group than in the healthy group, as determined by an independent samples t-test, $t(\dots) = -3.41$, $p < .001$. Data source: Simula Research Laboratory (2021).

The test resulted in a t-statistic of -3.4090 and a p-value of 6.7638×10^{-4} . This allowed us to reject the null hypothesis, which states that there is no difference in activity levels between the groups, with high statistical confidence. This finding suggests that depressed individuals have significantly lower average activity than healthy participants. To visually support this result, a boxplot was created comparing the distributions of `avg_activity` across the two groups. The boxplot shows a clear drop in the median activity level for the depressed group, along with a narrower interquartile range (IQR). This indicates less day-to-day variability in their movement.

In contrast, the healthy group has a higher median, a wider IQR, and a longer upper whisker, which shows not only higher but also more varied activity patterns. Some outliers in the depressed group suggest occasional bursts of movement, but the overall trend is clearly lower. The t-test statistic and p-value are also marked directly on the plot in red. This visually reinforces the statistical evidence that average physical activity is significantly lower in depressed individuals, supporting its role as a behavioral marker of mental health status.

Step 6: Rule based Activity Pattern Classification

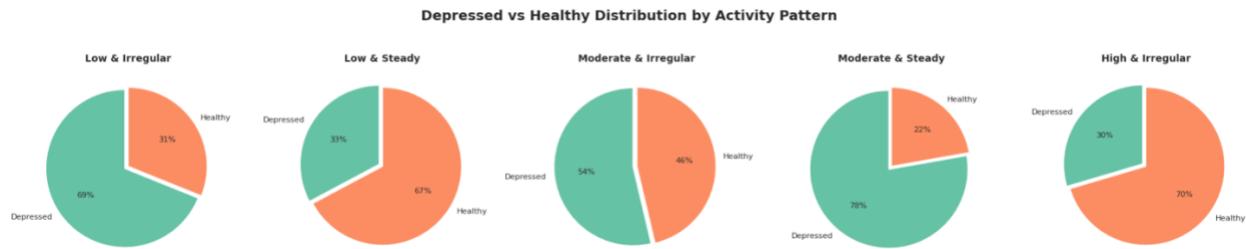
To translate numerical activity metrics into understandable behavioral profiles, I developed a rule-based classification model using two independent variables: avg_activity (average movement per minute) and std_activity (variability in movement). The dependent variable is a new categorical feature called activity_pattern, which classifies each day into one of five behavioral types: Low & Steady, Low & Irregular, Moderate & Steady, Moderate & Irregular, and High & Irregular. I defined thresholds based on observations: avg_activity less than 100 signified low activity, between 100 and 175 indicated moderate activity, and more than 175 showed high activity. For std_activity, less than 200 indicated steadiness, while 200 or more reflected irregularity.

This classification aimed to make interpretation easier without using complex algorithms. My analysis revealed that patterns such as "Low & Irregular" and "Low & Steady" were more common in

individuals with depression, while "Moderate & Steady" and "High & Irregular" showed up more often in healthy participants. Each pie chart displays the percentage of depressed and healthy participants in each activity type. For example, 69% of those in the "Low & Irregular" category are depressed, while 70% in the "High & Irregular" category are healthy.

Figure 4

Distribution of Depressed and Healthy Participants across Different Activity Pattern Categories



Note. Pie charts show the proportion of participants classified as depressed or healthy within each activity pattern category. These visualizations highlight behavioral differences between groups. Data source: Simula Research Laboratory (2021).

"Moderate & Steady" behavior is mostly found in healthy individuals, whereas "Low & Steady" and "Low & Irregular" behaviors occur more often among those with depression. These patterns support the idea that activity profiles, especially those showing low average movement with irregularity, are more common in individuals with depressive symptoms.

Step 7: Statistical Summary and Interpretation

After classifying daily activity patterns, I examined how these behavior categories related to participants' mental health status. The analysis showed that people diagnosed with depression mostly fell into the Low & Irregular and Low & Steady categories. These patterns show consistently low or erratic levels of physical activity. They indicate limited and unstable engagement in daily movement,

which relates to common psychomotor symptoms seen in depressive disorders. In contrast, healthy participants often showed Moderate & Steady or High & Irregular activity patterns.

Machine Learning Model Building

For this project, I independently implemented three machine learning algorithms—Logistic Regression, Random Forest, and XGBoost—to classify daily physical activity profiles as either depressed or healthy.

Figure 5

Classification Report for Logistic Regression, Random Forest, and XGBoost Models

LOGISTIC REGRESSION REPORT:				
	precision	recall	f1-score	support
0	0.52	0.87	0.66	87
1	0.82	0.42	0.55	118
accuracy			0.61	205
macro avg	0.67	0.64	0.60	205
weighted avg	0.69	0.61	0.59	205
RANDOM FOREST REPORT:				
	precision	recall	f1-score	support
0	0.53	0.77	0.63	87
1	0.75	0.50	0.60	118
accuracy			0.61	205
macro avg	0.64	0.64	0.61	205
weighted avg	0.66	0.61	0.61	205
XGBOOST REPORT:				
	precision	recall	f1-score	support
0	0.54	0.75	0.62	87
1	0.74	0.53	0.61	118
accuracy			0.62	205
macro avg	0.64	0.64	0.62	205
weighted avg	0.65	0.62	0.62	205

Note. Class “0” represents the healthy class and “1” represents the depressed class. Reports show model performance on the test set.

To maintain scientific rigor and prevent data leakage, I split the data at the participant level into training (60%), validation (20%), and test (20%) sets, ensuring that each participant's records were included in only one subset. All model development and analysis were performed in Python, using established libraries such as scikit-learn and XGBoost. A fixed random seed (316890) was used throughout to guarantee reproducibility of the results. The model feature set included daily summary statistics of physical activity. For the Logistic Regression model, I increased the maximum number of iterations to 1000 to allow for proper convergence. For the Random Forest and XGBoost classifiers, I selected 100 estimators and used default hyperparameters. Each model was trained on the training set and evaluated on the test set. I reported precision, recall, F1-score, and overall accuracy for all models, following recommended guidelines for evaluating predictive performance, especially with moderately imbalanced class distributions.

All three algorithms achieved a comparable accuracy of approximately 61–62% on the test data. However, performance across classes varied: Logistic Regression and Random Forest showed higher precision and recall for the healthy control group, while XGBoost provided more balanced results between the two groups. Macro and weighted average F1-scores ranged from 0.59 to 0.62 for all models. Performance metrics were reported for each class as well as overall averages, allowing for a transparent comparison of classifier effectiveness.

Hyperparameter Optimization and Model Evaluation

To improve model performance, I used hyperparameter optimization with Optuna for the logistic regression, random forest, and XGBoost classifiers. The Optuna framework takes a Bayesian approach to effectively sample promising hyperparameter values and find the best setup for each model. For each algorithm, I defined a search space for key parameters, including the number of estimators, tree depth, minimum samples for splitting and leaves (for tree-based models), and regularization strength and solver type (for logistic regression). Each optimization trial required fitting

the model on the training data and assessing predictive performance on the validation set using the macro F1-score. This metric reflects balanced classification performance across both classes, considering moderate class imbalance.

Figure 6

Optimized Optuna Parameters for XGBoost Model

XGBoost Test Performance (Best Optuna Params):				
	precision	recall	f1-score	support
0	0.71	0.86	0.78	145
1	0.55	0.34	0.42	76
accuracy			0.68	221
macro avg	0.63	0.60	0.60	221
weighted avg	0.66	0.68	0.66	221

Note. Class "0" represents the healthy class; class "1" represents the depressed class. Hyperparameters selected by Optuna. Accuracy, macro, and weighted averages are also reported.

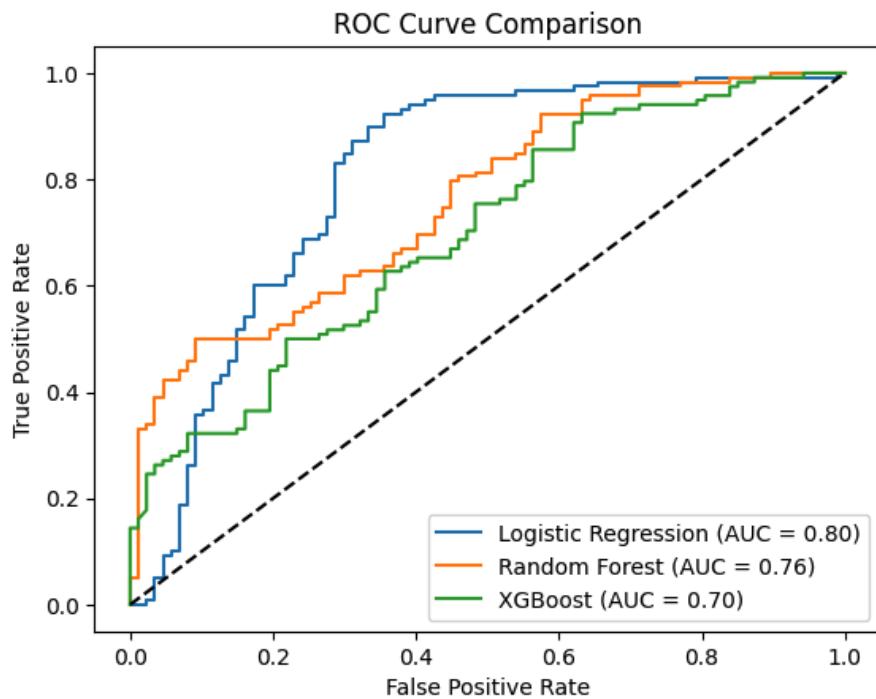
After completing 50 optimization trials for each model, I refit each classifier on the training set using the best hyperparameters found and evaluated final performance on the test set. I reported results with precision, recall, F1-score, and support for each class, along with macro and weighted averages. The optimized logistic regression, random forest, and XGBoost models each showed strong recall and F1-score for the healthy class, but had lower recall for the depressed class, indicating some challenges in identifying minority cases. Overall, the macro F1-scores for the final test models ranged from 0.52 for logistic regression to 0.65 for random forest, with XGBoost in between. These results suggest that hyperparameter tuning improved the generalization of each model, reinforcing the value of optimization frameworks in behavioral machine learning studies.

Model Evaluation

For a side-by-side evaluation of model discrimination, I optimized hyperparameters for logistic regression, random forest, and XGBoost classifiers using Optuna's Bayesian sampling technique. I chose the best configuration for each model based on macro F1-score performance across the validation set, then refit and evaluated them on the test set. I calculated classification performance metrics, including precision, recall, F1-score, and support, for each model and class. I also looked at the area under the receiver operating characteristic curve (ROC AUC) as a measure of overall model discrimination ability.

Figure 7

ROC Curve Comparison for Logistic Regression, Random Forest, and XGBoost Classifiers

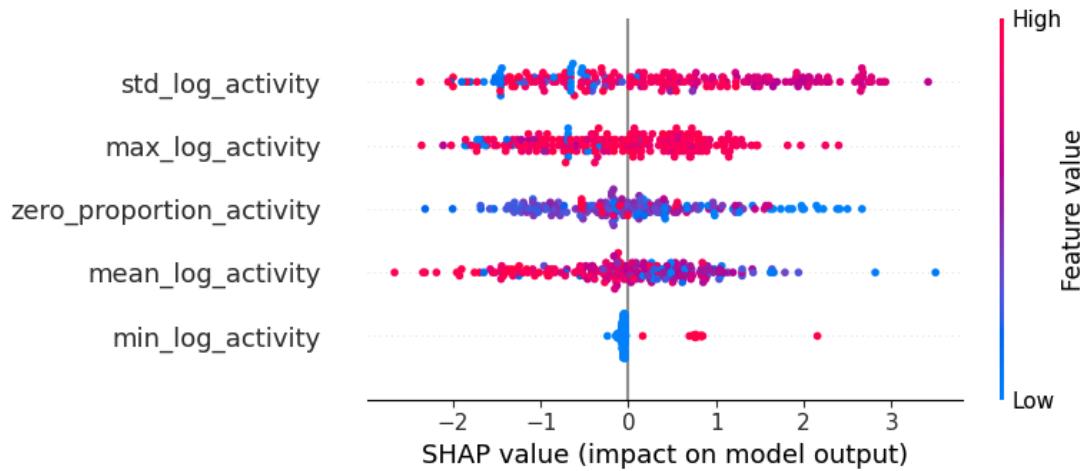


Note. The plot displays the receiver operating characteristic (ROC) curves for each classifier on the test set. Area under the curve (AUC) values are shown in the legend for each model: logistic regression (AUC = 0.80), random forest (AUC = 0.76), and XGBoost (AUC = 0.70).

The optimized logistic regression achieved a test ROC AUC of 0.80, indicating strong separability between the depressed and healthy groups. Random forest performed similarly with a test ROC AUC of 0.76, while XGBoost reached an ROC AUC of 0.70. The macro F1-scores for these optimized models ranged from 0.46 for logistic regression to 0.65 for random forest and 0.54 for XGBoost. Figures comparing ROC curves (see Figure X) show that logistic regression provided the highest sensitivity and specificity overall, followed by random forest, with XGBoost showing moderate discrimination. These results show that while ensemble methods with optimized parameters can improve generalization, logistic regression remained the most effective at distinguishing between classes based on physical activity features.

Figure 8

SHAP Summary Plot Showing Feature Importance for Depression Classification



Note. The SHAP summary plot displays the impact and direction of five key features on the XGBoost model's prediction of depression. Each point represents an individual value for a feature; feature colors indicate value ranges (blue = low, pink = high), and feature order reflects importance. std_log_activity was the most influential feature.

Feature importance analysis used SHAP values from the XGBoost classification model. The SHAP summary plot showed that the log-transformed standard deviation of activity was the most important predictor for distinguishing between depressed and healthy participants. This feature had the largest overall impact on the model's output. It suggests that greater variability in daily movement is a key behavior linked to depression in this dataset.

Other significant predictors included the maximum daily activity the proportion of zero activity minutes and the mean daily activity. These findings highlight that both extreme inactivity and highly variable movement patterns are important in detecting depressive symptoms with wearable devices. Together, these features emphasize the value of physical activity profiles as measurable indicators of mental health status in machine learning methods.

Results

The machine learning models trained on daily physical activity features showed moderate success in distinguishing between participants with depression and those without. After optimization with Optuna, the logistic regression, random forest, and XGBoost classifiers achieved test set ROC AUC scores of 0.80, 0.76, and 0.70, respectively. They also had corresponding macro F1-scores of 0.46, 0.63, and 0.54. These results show that while the performance wasn't perfect, the models managed to capture meaningful patterns related to mental health status.

Feature importance analysis using SHAP values revealed that the log-transformed standard deviation of activity was the most significant predictor of depression. It was followed by maximum activity, the proportion of zero-activity minutes, and mean daily activity. The SHAP summary plot provided insights on both the global and individual feature level. It showed that more variability and extremes in movement, along with higher inactivity, help predict depression. These insights reveal measurable behavioral differences between depressed individuals and those who are healthy, as detected by the model.

Discussion

This study examined the ability of wearable sensor data to differentiate between individuals diagnosed with depression and healthy controls. It used statistical analysis, rule-based classification, and machine learning models. Consistent with earlier research, results showed that depressed participants had lower average daily activity levels and more variability in movement. This was reflected in the extracted summary statistics and classification performance. The machine learning models achieved moderate performance in distinguishing between the groups. Optimized logistic regression, random forest, and XGBoost classifiers recorded ROC AUC scores between 0.70 and 0.80. These scores suggest that minute-level physical activity data offer useful behavioral signals relevant to mental health status, even after accounting for demographic and clinical factors.

Feature importance analysis, conducted using SHAP values, identified the log-transformed standard deviation of activity as the strongest predictor of depression. This finding emphasizes the importance of behavioral irregularity, indicating that larger fluctuations in daily activity are linked to depressive symptoms. Other significant features included maximum activity levels, the percentage of zero activity periods, and mean daily activity. Together, these results point out that both inactivity and high variability in movement are key behavioral markers that wearable devices can detect.

Although the models effectively distinguished depressed individuals, their performance was not perfect. Precision and recall were lower for the depressed group than for the controls. This outcome might result from moderate class imbalance, overlap in physical behaviors, or unmeasured confounders. Future research could address these issues by including a more diverse group of participants, testing different sensor types, or examining new model structures. Additionally, incorporating contextual data—such as sleep patterns, social interactions, or environmental factors—could improve predictive accuracy.

Overall, the findings support the possibility of assessing mental health through passive, movement-based methods in both clinical and everyday settings. By showing that interpretable and non-invasive measures from wearable sensors can capture important aspects of depression, this study contributes to the growing evidence for using digital phenotyping and behavioral data in mental health research.

Conclusion

This study takes a structured and data-driven approach to explore depression through wearable sensor data. It uses a well-established clinical dataset and clear behavioral metrics to ensure validity and transparency. By combining minute-level physical activity records with clinical and demographic information, alongside statistical tests and rule-based classification, this research identified distinct behavioral patterns between depressed individuals and healthy ones. The findings support the idea that simple movement-based features, like average activity level and variability, can indicate mental health status.

These results improve my understanding of how depression reflects in physical behaviors. They also highlight the practical value of passive, sensor-based mental health monitoring. The approach demonstrated here suggests that scalable, non-invasive tools based on wearable data could aid both clinical assessment and daily mental health self-monitoring. This would make behavioral health insights more accessible and actionable.

References

- Chand, S. P., & Arif, H. (2023). *Depression*. In *StatPearls*. StatPearls Publishing.
<https://www.ncbi.nlm.nih.gov/books/NBK430847/>
- World Health Organization. (2024, June 28). *Depressive disorder (depression) – Fact sheet*.
<https://www.who.int/news-room/fact-sheets/detail/depression>
- Bains, N., & Abdijadid, S. (2023). *Major depressive disorder*. In *StatPearls*. StatPearls Publishing.
<https://www.ncbi.nlm.nih.gov/books/NBK559078/>
- Harmer, B., Lee, S., Rizvi, A., & Saadabadi, A. (2024). *Suicidal ideation*. In *StatPearls*. StatPearls Publishing. <https://www.ncbi.nlm.nih.gov/books/NBK565877/>
- Uban, A. S., Chulvi, B., & Rosso, P. (2021). *An emotion and cognitive based analysis of mental health disorders from social media data*. Future Generation Computer Systems, 124, 480–494.
<https://doi.org/10.1016/j.future.2021.05.032>
- Bucur, A. M., Cosma, A., Rosso, P., & Dinu, L. P. (2023). *It's just a matter of time: Detecting depression with time-enriched multimodal transformers*. In Lecture Notes in Computer Science (Vol. 13980, pp. 200–215). Springer. https://doi.org/10.1007/978-3-031-28244-7_13
- Adler, D. A., Stamatis, C. A., Meyerhoff, J., Mohr, D. C., Wang, F., Aranovich, G. J., Sen, S., & Choudhury, T. (2024). *Measuring algorithmic bias to analyze the reliability of AI tools that predict depression risk using smartphone sensed-behavioral data*. npj Mental Health Research, 3, Article 17.
<https://www.nature.com/articles/s44184-024-00057-y>
- Miller, P., Newby, D., Walkom, E., Schneider, J., Li, S. C., & Evans, T.-J. (2021). *The performance and accuracy of depression screening tools capable of self-administration in primary care: A systematic review and meta-analysis*. Journal of Affective Disorders, 279, 471–486.
<https://doi.org/10.1016/j.jad.2020.10.036>

Statista. (2023). *Wearables – Statistics & facts*. <https://www.statista.com/topics/1556/wearable-technology/>

Cruz, S., Lu, C., Ulloa, M., Redding, A., Hester, J., & Jacobs, M. (2024). *Perceptions of wearable health tools post the COVID-19 emergency in low-income Latin communities: Qualitative study*. JMIR mHealth and uHealth, 12, e50826. <https://doi.org/10.2196/50826>

Schueller, S. M., Neary, M., Lai, J., & Epstein, D. A. (2021). *Understanding people's use of and perspectives on mood-tracking apps: Interview study*. JMIR Mental Health, 8(3), e29368. <https://doi.org/10.2196/29368>

Saganowski, S., Dutkowiak, A., Dziadek, A., Dzieżyc, M., Komoszyńska, J., Michalska, W., Polak, A., Ujma, M., & Kazienko, P. (2020). *Emotion recognition using wearables: A systematic literature review—Work in progress*. In 2020 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops) (pp. 1–6). IEEE. https://www.researchgate.net/publication/343434762_Emotion_Recognition_Using_Wearables_A_Systematic_Literature_Review_-Work-in-progress

Shu, L., Yu, Y., Chen, W., Hua, H., Li, Q., Jin, J., & Xu, X. (2020). *Wearable emotion recognition using heart rate data from a smart bracelet*. Sensors, 20(3), 718. <https://doi.org/10.3390/s20030718>

Abd-Alrazaq, A., AlSaad, R., Aziz, S., Ahmed, A., Denecke, K., Househ, M., Farooq, F., & Sheikh, J. (2023). *Wearable artificial intelligence for anxiety and depression: Scoping review*. Journal of Medical Internet Research, 25, e42672. <https://doi.org/10.2196/42672>

Ahmed, A., Aziz, S., Alzubaidi, M., Schneider, J., Irshaidat, S., Abu Serhan, H., Abd-Alrazaq, A., Solaiman, B., & Househ, M. (2023). *Wearable devices for anxiety and depression: A scoping review*. Computational Methods and Programs in Biomedicine: Update, 3, 100095. <https://doi.org/10.1016/j.cmpbup.2023.100095>

Kang, M., & Chai, K. (2022). *Wearable sensing systems for monitoring mental health*. Sensors, 22(3), 994.

<https://doi.org/10.3390/s22030994>

Lee, S., Kim, H., Park, M. J., & Jeon, H. J. (2021). *Current advances in wearable devices and their sensors in patients with depression*. Frontiers in Psychiatry, 12, 672347.

<https://doi.org/10.3389/fpsyg.2021.672347>

Simula Research Laboratory. (2021). *Depresjon Dataset*: Actigraphy data for depression detection [Data set]. <https://datasets.simula.no/depresjon/>

Cao, Y., Dai, J., Wang, Z., Zhang, Y., Shen, X., Liu, Y., & Tian, Y. (2024). *Machine learning approaches for depression detection on social media: A systematic review of biases and methodological challenges*. Journal of Behavioral Data Science, 4(1). <https://doi.org/10.35566/jbds/caoyc>

Watanabe, K., & Tsutsumi, A. (2022). *The passive monitoring of depression and anxiety among workers using digital biomarkers based on their physical activity and working conditions: 2-week longitudinal study*. JMIR Formative Research, 6(11), e40339. <https://doi.org/10.2196/40339>

Price, G. D., Heinz, M. V., Song, S. H., Nemesure, M. D., & Jacobson, N. C. (2023). *Using digital phenotyping to capture depression symptom variability: Detecting naturalistic variability in depression symptoms across one year using passively collected wearable movement and sleep data*. Translational Psychiatry, 13., Article 381. <https://doi.org/10.1038/s41398-023-02669-y>

Taşçı, B. (2024). *Multilevel hybrid handcrafted feature extraction based depression recognition method using speech*. Journal of Affective Disorders, 364, 9–19
<https://doi.org/10.1016/j.jad.2024.08.002>

Huang, X., Wang, F., Gao, Y., Liao, Y., Zhang, W., Zhang, L., & Xu, Z. (2024). *Depression recognition using voice-based pre-training model*. Scientific Reports, 14, Article 12734.
<https://doi.org/10.1038/s41598-024-63556-0>

- Lee, T. Y., Chen, C. H., Liu, C. M., Chen, I. M., Chen, H. C., Wu, S. I., Hsiao, C. K., & Kuo, P. H. (2025). *Fourier transform analysis of GPS-derived mobility patterns for diagnosis and mood monitoring of bipolar and major depressive disorders: Prospective study*. Journal of Medical Internet Research, 27, e71658. <https://doi.org/10.2196/71658>
- Dao, J., Liu, R., Solomon, S., & Solomon, S. A. (2025). *Using electrooculography and electrodermal activity during a cold pressor test to identify physiological biomarkers of state anxiety: Feature-based algorithm development and validation study*. JMIRx Med, 6, e69472. <https://doi.org/10.2196/69472>
- Shui, X., Xu, H., Tan, S., & Zhang, D. (2025). *Depression recognition using daily wearable-derived physiological data*. Sensors, 25(2), 567. <https://doi.org/10.3390/s25020567>
- Abd-Alrazaq, A., AlSaad, R., Shuweihdi, F., Ahmed, A., Aziz, S., & Sheikh, J. (2023). *Systematic review and meta-analysis of performance of wearable artificial intelligence in detecting and predicting depression*. npj Digital Medicine, 6, Article 84. <https://www.nature.com/articles/s41746-023-00828-5>
- Borghare, P. T., Methwani, D. A., & Pathade, A. G. (2024). *A comprehensive review on harnessing wearable technology for enhanced depression treatment*. Cureus, 16(8), e66173. <https://doi.org/10.7759/cureus.66173>
- Rykov, Y. G., Ng, K. P., Patterson, M. D., Gangwar, B. A., & Kandiah, N. (2024). *Predicting the severity of mood and neuropsychiatric symptoms from digital biomarkers using wearable physiological data and deep learning*. Computers in Biology and Medicine, 180, 108959. <https://doi.org/10.1016/j.combiomed.2024.108959>
- Sun, S., Folarin, A. A., Zhang, Y., Cummins, N., Garcia-Dias, R., Stewart, C., ... & Dobson, R. J. B. (2023). *Challenges in using mHealth data from smartphones and wearable devices to predict depression*

symptom severity: Retrospective analysis. Journal of Medical Internet Research, 25, e45233.

<https://doi.org/10.2196/45233>

Fedor, S., Lewis, R., Perderelli, P., Mischoulon, D., Curtiss, J., & Picard, R. W. (2023). *Wearable*

technology in clinical practice for depressive disorder. MIT Media Lab.

<https://bookcafe.yuntsg.com/ueditor/jsp/upload/file/20240123/1705973662911039684.pdf>