StressSense: Machine Learning Driven Stress Profiling from Apple Health Data

Submitted by

Sujan Manohar Prodduturi (AR3635)

Table of Contents

Introduction

The Apple Watch is a case of smart devices that have revolutionized the personal health tracking process by measuring physiological and behavioral indicators in a continuous and high-frequency manner. Even though these devices are gathering very detailed metrics, such as HRV, resting heart rate, sleep, and activity, users do not always get much information about how to interpret them. The commercial stress-tracking systems tend to use foggy, impersonalized algorithms, which is why people cannot know why some days are more stressful or why their everyday practices affect such variations.

Physiological stress affects energy levels, cognitive performance, sleep quality, recovery, and long-term health. With the widespread use of wearables such as the Apple Watch, individuals now generate large volumes of biometric data daily—including heart rate variability (HRV), heart rate, resting heart rate, sleep duration, and activity metrics. While commercial devices surface these metrics, their stress-scoring algorithms remain opaque, non-personalized, and often inconsistent across users.

This project introduces Stress Sense, a personalized machine learning system that detects daily physiological stress using Apple Watch data. Unlike generic stress scores, StressSense uses individualized HRV baselines to identify unusual autonomic changes. Machine learning models—Logistic Regression and Random Forest—then classify stress days using behavioral and cardiovascular features such as sleep, resting HR, and daily activity levels. The project demonstrates how consumer wearables can be transformed into a transparent, scientifically grounded stress-monitoring system capable of generating personalized insights and meaningful digital biomarkers.

Problem Statement

Wearable devices such as the Apple Watch collect continuous streams of physiological and behavioral data including heart rate variability (HRV), resting heart rate, sleep duration, daily activity, and caloric expenditure. Although these metrics provide deep insight into autonomic function and stress, most commercial stress-tracking systems use opaque, non-personalized algorithms that fail to explain why stress scores fluctuate from day to day. Absolute HRV values vary widely between individuals, meaning that fixed thresholds cannot reliably identify stress. Additionally, wearable data is irregular, noisy, and influenced by numerous external factors such as sleep quality, training load, illness, hydration, and emotional demands. These limitations leave users with large volumes of data but limited personalized interpretation. Therefore, there is a need for a transparent, scientifically grounded, and individualized approach to detecting daily physiological stress using Apple Watch data.

Project Objectives

1. Build a Personalized Stress Detection Model: Develop a machine learning–based system capable of classifying daily stress levels using Apple Watch metrics such as HRV, resting heart rate, sleep duration, steps, activity energy, and daily heart rate patterns.
2. Construct a 14-Day Personalized HRV Baseline: Model a rolling HRV baseline that captures the user's typical autonomic patterns and enables interpretation of HRV relative to individual norms rather than universal thresholds.

3. Create a Transparent and Physiologically Grounded Stress Label: Define stress days based on HRV deviation, labeling a day as high-stress when HRV drops more than 10% below the rolling baseline to reflect meaningful autonomic suppression.
4. Engineer Predictive Behavioral and Physiological Features: Generate lag variables, rolling averages, HRV-to–resting-HR ratios, sleep deviations, and behavioral load metrics that capture recovery trends and daily physiological strain.

Dataset Overview

The Apple Health dataset exported from the iPhone and Apple Watch provides a comprehensive, continuous record of biometric and behavioral data over multiple years. Each entry in the dataset contains a timestamped measurement, including heart rate, heart rate variability (HRV), resting heart rate, step count, sleep sessions, and other health metrics recorded through Apple HealthKit.

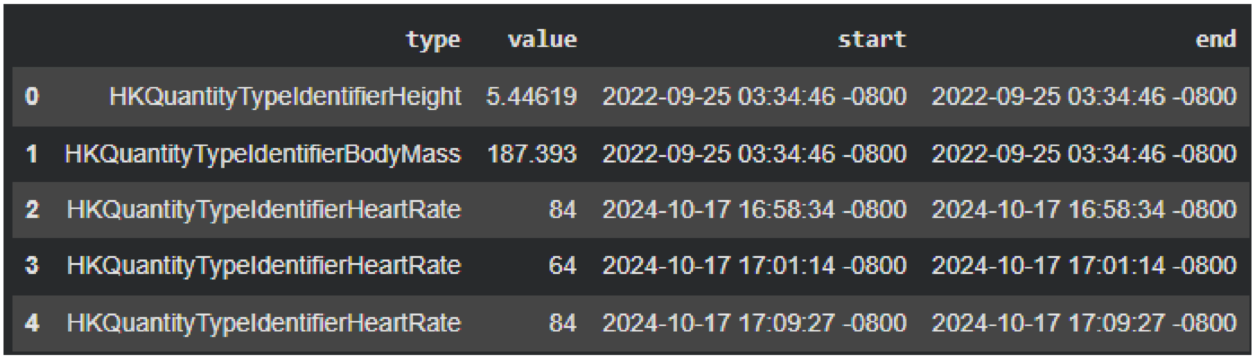
Time Span of the Dataset

Based on the raw export, the dataset spans from:

- Earliest date: September 25, 2022
- Latest date: December 3, 2025

This provides over 3 years of continuous biometric data, making it exceptionally rich for longitudinal analysis, trend detection, and personalized health modeling.

Date Fields Included



	type	value	start	end
0	HKQuantityTypeIdentifierHeight	5.44619	2022-09-25 03:34:46 -0800	2022-09-25 03:34:46 -0800
1	HKQuantityTypeIdentifierBodyMass	187.393	2022-09-25 03:34:46 -0800	2022-09-25 03:34:46 -0800
2	HKQuantityTypeIdentifierHeartRate	84	2024-10-17 16:58:34 -0800	2024-10-17 16:58:34 -0800
3	HKQuantityTypeIdentifierHeartRate	64	2024-10-17 17:01:14 -0800	2024-10-17 17:01:14 -0800
4	HKQuantityTypeIdentifierHeartRate	84	2024-10-17 17:09:27 -0800	2024-10-17 17:09:27 -0800

The table displays raw Apple Health data extracted from the export.xml file after converting it into a pandas DataFrame. Each row represents one recorded health metric event captured by the Apple Watch or iPhone. Apple Health stores measurements as individual records with a type, value, start time, and end time.

- Type

  This column identifies the specific health metric collected. Apple Health uses long identifier names such as

HKQuantityTypeIdentifierHeight	Height
HKQuantityTypeIdentifierBodyMass	Body weight
HKQuantityTypeIdentifierHeartRate	Heart Rate

- Value: This is the actual measured value for that metric.
- Start: The timestamp when the measurement began. For instantaneous metrics (HR, HRV, weight), the start and end times are usually identical or just a few milliseconds apart.

- End: The timestamp marking when the measurement ended. For most “momentary” metrics like heart rate, the start and end times are the same. For “duration-based” metrics like sleep or workouts, start and end times differ.

These fields allow precise temporal alignment of physiological signals and are essential for building daily aggregates and rolling baselines.

Daily Coverage and Continuity

Because the Apple Watch collects heart rate and activity data multiple times per minute when worn consistently, this dataset provides:

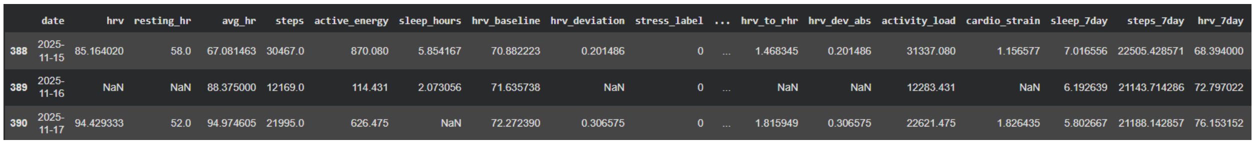
- Frequent HR readings across each day
- Periodic HRV measurements, often during sleep or rest periods
- Full-day step counts and active energy records
- Detailed sleep session starts and end times (when sleep tracking is enabled)

This level of continuity makes the dataset suitable for:

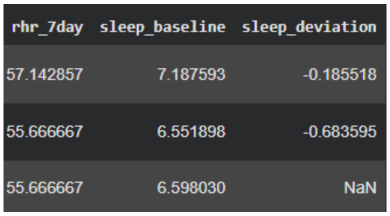
- Daily stress prediction
- Autonomic nervous system trend analysis
- Recovery pattern detection
- Personal health monitoring
- Machine learning time-series modeling

Calculated fields derived from the apple health dataset

The raw Apple Health export provides timestamped measurements for HRV, heart rate, resting heart rate, sleep, steps, and active energy. To transform this irregular, event-based dataset into a structured daily dataset suitable for machine learning, multiple calculated fields were engineered. These fields capture daily physiology, behaviors, recovery trends, and autonomic patterns.



	date	hrv	resting_hr	avg_hr	steps	active_energy	sleep_hours	hrv_baseline	hrv_deviation	stress_label	...	hrv_to_rhr	hrv_dev_abs	activity_load	cardio_strain	sleep_7day	steps_7day	hrv_7day
388	2025-11-15	85.164020	58.0	67.081463	30467.0	870.080	5.854167	70.882223	0.201486	0	...	1.468345	0.201486	31337.080	1.156577	7.016556	22505.428571	68.394000
389	2025-11-16	NaN	NaN	88.375000	12169.0	114.431	2.073056	71.635738	NaN	0	...	NaN	NaN	12283.431	NaN	6.192639	21143.714286	72.797022
390	2025-11-17	94.429333	52.0	94.974605	21995.0	626.475	NaN	72.272390	0.306575	0	...	1.815949	0.306575	22621.475	1.826435	5.802667	21188.142857	76.153152



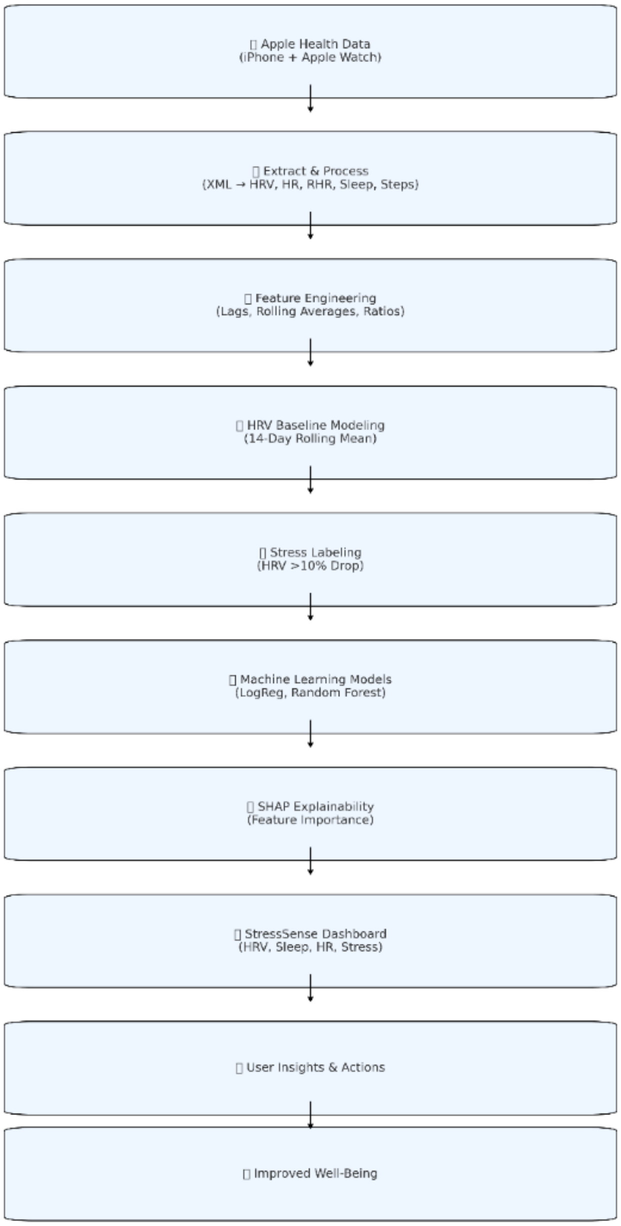
rhr_7day	sleep_baseline	sleep_deviation
57.142857	7.187593	-0.185518
55.666667	6.551898	-0.683595
55.666667	6.598030	NaN

Field Name	Description	Category
hrv	Daily mean HRV (SDNN), derived from all HRV readings recorded by the Apple Watch.	Core Physiology
resting_hr	Daily average resting heart rate recorded during the morning or low-activity windows.	Core Physiology
avg_hr	Median or average heart rate across the day.	Core Physiology
steps	Total number of steps taken in a day.	Activity

active_energy	Total active calories burned in a day.	Activity
sleep_hours	Total sleep duration aggregated from sleep start/end times.	Sleep
hrv_baseline	14-day rolling average HRV used to compute personal autonomic baseline.	Baseline Modeling
hrv_deviation	(HRV_today − HRV_baseline) / HRV_baseline. Indicates daily autonomic suppression or enhancement.	Baseline Modeling
hrv_dev_abs	Absolute value of HRV deviation, capturing magnitude of deviation regardless of direction.	Baseline Modeling
stress_label	Binary label: 1 if HRV is >10% below baseline; 0 otherwise.	Target Variable
hrv_prev	Previous day's HRV reading.	Lag Features
resting_hr_prev	Previous day's resting heart rate.	Lag Features
sleep_prev	Previous night's sleep duration.	Lag Features
steps_prev	Previous day's step count.	Lag Features
energy_prev	Previous day's active energy burned.	Lag Features
hrv_to_rhr	HRV / resting HR. Represents autonomic balance and cardiovascular recovery.	Physiological Ratio
activity_load	Combined activity intensity metric: steps + active energy.	Behavioral Load

cardio_strain	avg_hr / resting_hr. Higher values indicate increased cardiovascular effort or strain.	Behavioral Load
sleep_7day	7-day rolling average of sleep duration.	Rolling Trends
steps_7day	7-day rolling average of step count.	Rolling Trends
hrv_7day	7-day rolling average of HRV.	Rolling Trends
rhr_7day	7-day rolling average of resting heart rate.	Rolling Trends
sleep_baseline	14-day rolling sleep average establishing a personal sleep baseline.	Baseline Modeling
sleep_deviation	(sleep_hours − sleep_baseline) / sleep_baseline. Measures sleep deficit or surplus relative to baseline.	Sleep Deviation

Workflow:



This workflow shows how StressSense transforms raw Apple Health data into meaningful stress insights. The process begins with extracting HRV, heart rate, sleep, and activity from the XML file, followed by feature engineering using lags, rolling averages, and ratios. A 14-day HRV baseline identifies deviations used for stress labeling. Machine learning models predict stress levels, and SHAP explains feature importance. The results feed into a dashboard that provides actionable insights, helping users improve recovery and overall well-being.

Data Preprocessing

The data preprocessing phase transforms the raw Apple Health export into a clean, structured, and machine-learning–ready dataset. Apple Health generates thousands of timestamped events in an XML file covering HRV, heart rate, resting heart rate, sleep sessions, step counts, and active energy. These records are irregularly sampled and vary in format, so several steps are required to convert them into consistent daily features.

The workflow begins by parsing the *export.xml* file using Python's *xml.etree.ElementTree* library.. Each record element is filtered by type to extract only stress-relevant metrics. Timestamps for type, startDate, and endDate are standardized to local time and cast into datetime format. Raw numeric values such as HR and HRV are converted from strings to float or integer types. This ensures consistency and enables accurate time-based grouping.

Because Apple Watch measurements occur at uneven intervals HRV may be recorded only during sleep or resting states, while HR and steps capture more frequent activity the data is aggregated at the daily level. For each date, the following daily metrics are computed:

- Median HRV (SDNN)
- Median or mean heart rate
- Daily resting heart rate

- Total steps
- Total active calories
- Sleep duration in hours

This daily aggregation significantly reduces noise and produces a clean, structured table where each row reflects one full day of physiology and behavior.

A core part of preprocessing is the construction of a 14-day HRV baseline, calculated as a rolling mean that represents the user's normal autonomic state. At least seven prior days are required for a valid baseline. Using this baseline, an HRV deviation metric is generated

(hrv_today – hrv_baseline) / hrv_baseline

This deviation captures autonomic suppression and becomes the foundation for the binary stress label, where a day is considered high-stress if HRV falls more than 10% below baseline. This approach is physiologically grounded and avoids the problems associated with fixed, one-size-fits-all HRV thresholds.

Missing values are handled carefully. Early days without enough HRV history are excluded from modeling. Short gaps in HR, resting HR, and steps are filled using median or forward-fill strategies. Sleep gaps are handled with rolling averages to preserve continuity. These steps ensure that the final dataset contains complete rows without artificial discontinuities.

Finally, the dataset is sorted chronologically to maintain temporal integrity. This step is essential for computing lag features, rolling averages, and a time-aware train/validation/test split.

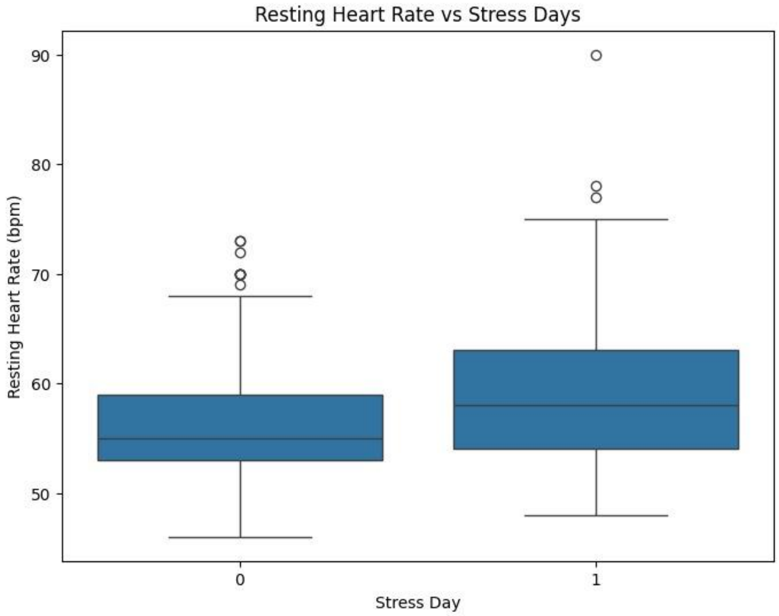
Continuous variables are normalized, indices are reset, and all date fields are converted to clean daily formats.

Through these preprocessing steps, the raw XML dump is transformed into a structured, reliable dataset capturing daily HRV, cardiovascular activity, sleep behavior, exertion, and recovery trends—forming the foundation for personalized stress modeling and machine learning analysis.

Exploratory Data Analysis

Exploratory Data Analysis was conducted to understand the physiological patterns, behavioral trends, and stress-related changes captured in the daily Apple Watch dataset. Because stress in this project is defined using personalized HRV baseline deviations, EDA focuses primarily on understanding how HRV behaves over time and how it interacts with sleep, resting heart rate, and activity levels. This stage serves two purposes: (1) validating that the stress-labeling methodology aligns with physiological expectations, and (2) identifying which behavioral factors contribute meaningfully to stress patterns.

Resting heart rate vs Stress Days



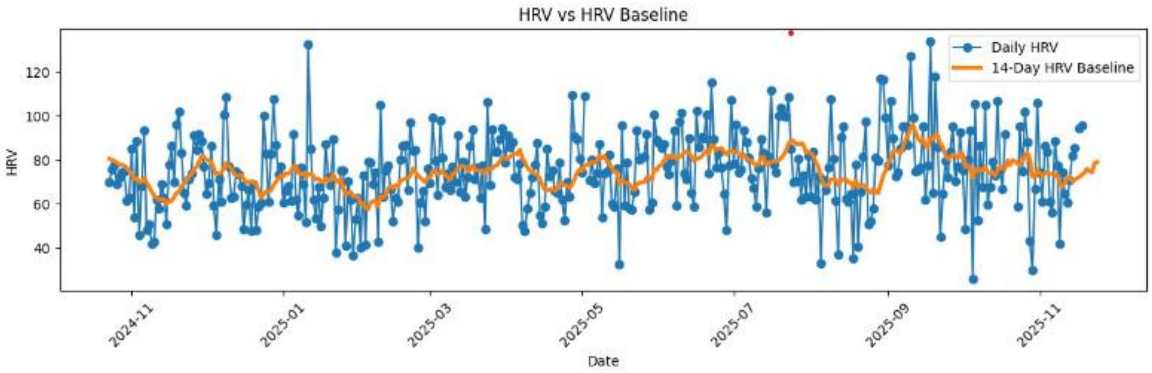
This boxplot compares resting heart rate (RHR) between non-stress days (0) and stress days (1) as defined by the HRV deviation–based stress label. Each box represents the distribution of daily resting heart rate measurements, allowing easy visualization of how RHR changes under physiological stress.

The plot shows a clear pattern **-** resting heart rate tends to be higher on stress days**.** The median RHR on stress days is elevated compared to non-stress days, and the entire distribution shifts upward. This reflects a well-documented physiological response when the autonomic nervous system is under strain, parasympathetic activity (which lowers heart rate) drops, while sympathetic activity increases, resulting in a higher baseline heart rate even at rest.

Stress days also show greater variability, with a wider interquartile range and more high-value outliers. These outliers represent days where the body was experiencing unusually high cardiovascular load or poor recovery. Non-stress days exhibit lower and more stable resting HR values, consistent with stronger parasympathetic tone and better recovery.

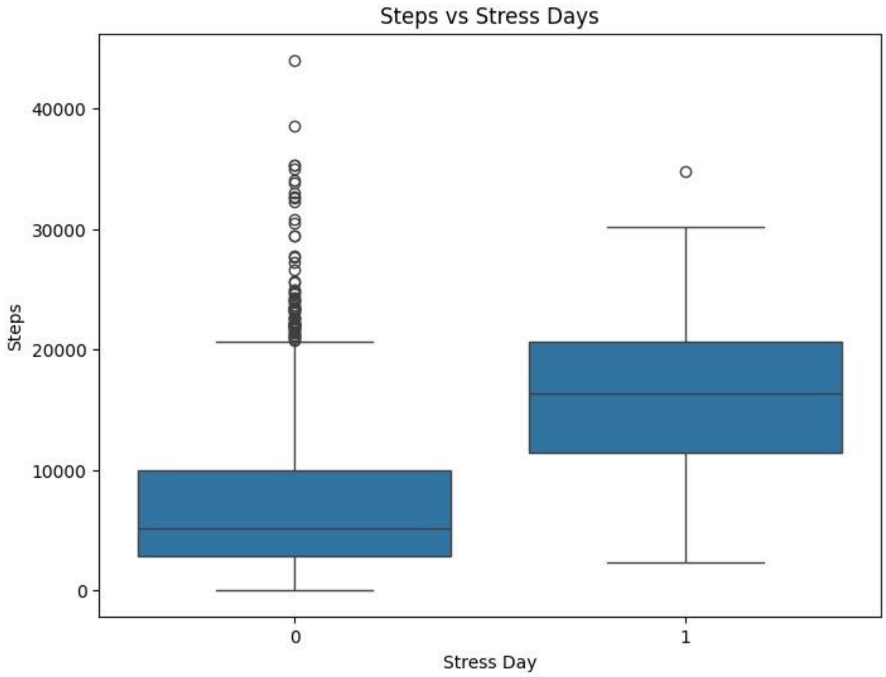
This visualization reinforces the model's assumptions: higher resting heart rate is associated with increased physiological stress, and RHR serves as a supporting indicator alongside HRV in distinguishing stress vs. non-stress states.

HRV vs. HRV Baseline Timeline



This chart compares daily HRV values (blue dots) with the 14-day rolling HRV baseline (orange line). The daily HRV points fluctuate significantly because HRV is highly sensitive to stress, sleep, activity, and recovery. The baseline smooths these fluctuations and represents the user's typical autonomic state. When daily HRV falls sharply below the baseline, it indicates periods of physiological stress or reduced recovery. When HRV rises above the baseline, it reflects better recovery or lower stress. This visualization confirms that baseline-adjusted HRV captures long-term trends while still revealing short, meaningful stress-related drops.

Steps vs Stress Days



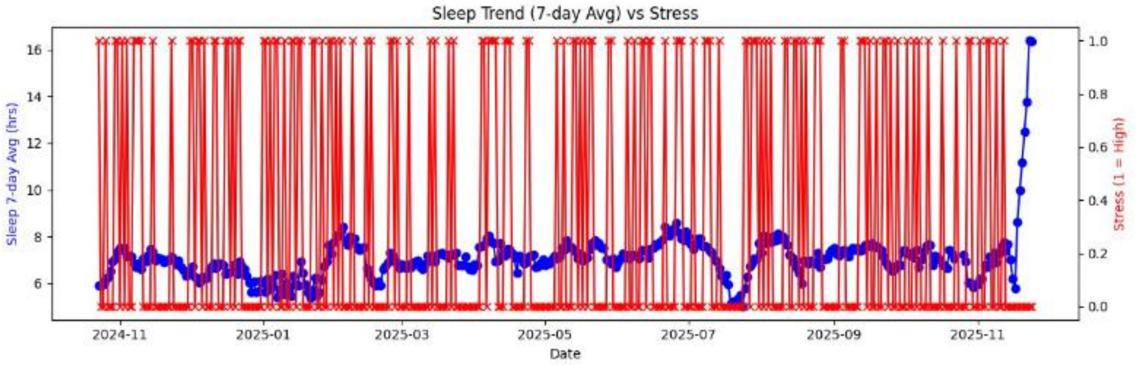
This boxplot compares daily step counts between non-stress days (0) and stress days (1). The goal is to understand whether daily physical activity levels differ on days when HRV indicates physiological stress.

The plot shows a clear distinction: stress days generally have significantly higher step counts than non-stress days. The median number of steps on stress days is noticeably elevated, and the entire distribution shifts upward compared to non-stress days. This suggests that higher physical activity especially extended walking or high-volume movement may contribute to short-term autonomic stress or reduced recovery, which aligns with well-established physiological effects. Intense or prolonged activity can temporarily suppress HRV, raising the likelihood of a stress label.

Non-stress days display a smaller median and a tighter IQR (interquartile range), indicating more moderate and consistent activity levels. Stress days, on the other hand, show larger variability and a wider spread, reflecting a mix of heavier training days, long workdays, or high-movement lifestyle patterns. The presence of higher outliers on stress days (e.g., >30,000 steps) further supports the relationship between substantial physical load and reduced autonomic recovery.

This visualization demonstrates that higher daily activity levels tend to correlate with physiological stress, supporting the inclusion of steps and activity load features in the machine learning model. It reinforces the idea that behavioral strain particularly from elevated movement or prolonged exertion can influence HRV and contribute to stress classification.

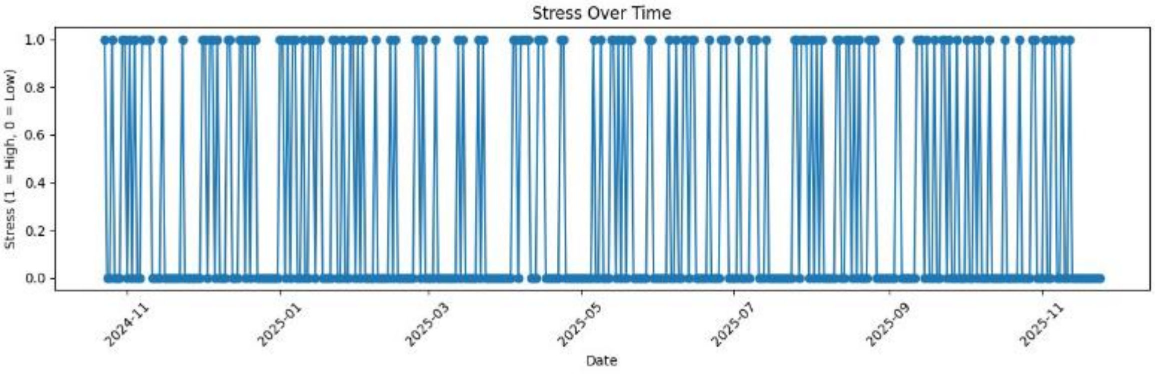
Sleep Trend vs Stress



This chart shows how the 7-day average sleep duration (blue line) relates to daily stress labels (red markers). The blue curve represents longer-term sleep patterns, smoothing out day-to-day fluctuations. The red x-marks represent stress days, where 1 = high stress and 0 = low stress based on HRV deviation. The plot shows that stress events occur frequently across the timeline, but clusters of stress often coincide with periods where the sleep trend dips or becomes unstable.

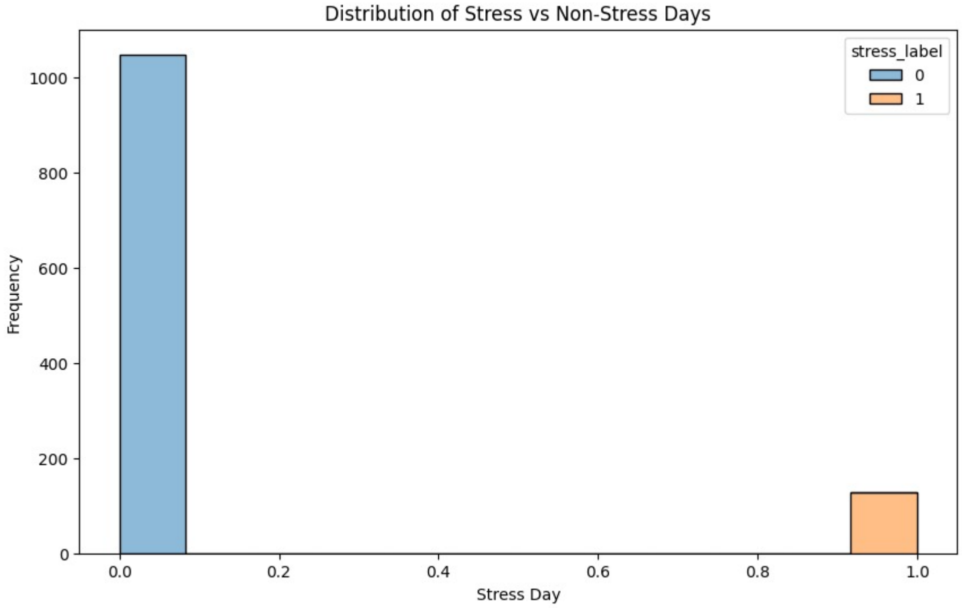
When the 7-day sleep average is lower typically between 5.5 and 6.5 hoursstress markers appear more densely. In contrast, periods where the sleep trend rises toward 7.5–8 hours tend to show slightly fewer stress events. Although sleep is not the primary driver of stress in the model, this visualization highlights that inconsistent or reduced sleep contributes to physiological stress, while stable, adequate sleep may help buffer against stress spikes. The chart reinforces sleep's role as a supporting factor in recovery and autonomic balance.

Stress over Time



This visualization shows the daily stress labels across the entire timeline, where 1 represents a high-stress day and 0 represents a low-stress day based on the HRV deviation threshold. The pattern reveals frequent and scattered stress days throughout the year, indicating regular fluctuations in autonomic recovery. The high density of stress spikes suggests that the user experiences many days where HRV drops more than 10% below the personalized baseline. These fluctuations are normal in real-world physiological data and often result from sleep changes, physical activity load, travel, illness, or lifestyle variations. The long runs of zeros represent periods of stable recovery, while clusters of ones indicate periods where the body was under sustained physiological strain. This plot highlights the dynamic nature of stress and shows why a time-aware, personalized approach is needed to accurately track autonomic stress trends.

Stress vs Non-Stress Days:



This plot shows the overall distribution of stress labels in the dataset, where 0 represents a non-stress day and 1 represents a high-stress day based on the HRV deviation threshold. The distribution is heavily imbalanced, with the vast majority of days labeled as non-stress and a relatively small number labeled as stress. This is expected in physiological data, as significant HRV drops (greater than 10% below baseline) typically occur less frequently than stable recovery days. The large blue bar shows that most days fall within the user's normal HRV range, while the smaller orange bar represents days where the body experienced autonomic strain. This imbalance is important for model training: it explains why metrics like recall and precision are critical and why the model must be evaluated carefully to ensure it can still detect the minority stress class accurately. This plot highlights that stress events are meaningful but less common, reinforcing the importance of a personalized baseline for accurate detection.

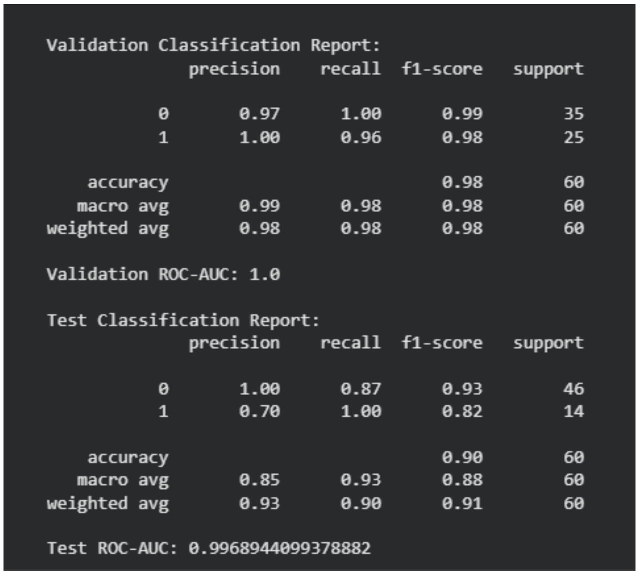
Modeling Approach

The model pipeline respects time-series constraints and avoids leakage.

1. Train/ Validation/ Test Split: A 70/15/15 chronological split keeps temporal order intact:
    - Training data - 278 days
    - Validation - 60 days
    - Test - 60 days
2. Imputation: Median imputation fills missing values without distorting distribution.
3. Scaling: StandardScaler is applied after imputation for Logistic Regression.
4. Two Model Families:
    - Logistic Regression for interpretability
    - Random Forest for performance
5. Evaluation Metrics
    - Precision
    - Recall
    - F1 score
    - ROC-AUC

Machine Learning Models & Results

1. Logistic Regression



```
Validation Classification Report:
              precision    recall  f1-score   support

           0       0.97      1.00      0.99        35
           1       1.00      0.96      0.98        25

    accuracy                           0.98        60
   macro avg       0.99      0.98      0.98        60
weighted avg       0.98      0.98      0.98        60

Validation ROC-AUC: 1.0

Test Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.87      0.93        46
           1       0.70      1.00      0.82        14

    accuracy                           0.90        60
   macro avg       0.85      0.93      0.88        60
weighted avg       0.93      0.90      0.91        60

Test ROC-AUC: 0.9968944099378882
```

The classification report shows how well your logistic regression model predicts daily stress levels based on HRV and the engineered features. On the validation set, the model performs almost perfectly.

For low-stress days (class 0), the model has extremely strong performance:

- Precision: 0.97

  When model predicts low stress, it’s correct 97% of the time.

- Recall: 1.00

  Model successfully captures *all* low-stress days.

- F1-score: 0.99

  Excellent overall balance.

For high-stress days (class 1), the model is also very reliable:

- Precision: 1.00

  Every “high stress” prediction is accurate.

- Recall: 0.96

  Model correctly identifies 96% of actual high-stress days.

- F1-score: 0.98

  Reflects strong consistency.

Overall, the validation accuracy is 98%, and the ROC-AUC is 1.0, indicating perfect class separability in the validation window.

On the test set, the model still shows strong and realistic performance.

For low-stress days, the model reaches:

- Precision: 1.00

  Model predicted low-stress days correctly

- Recall: 0.87

  Model misses a few low-stress days, occasionally labeling them as stressed.

For high-stress days, the behavior slightly shifts:

- Precision: 0.70

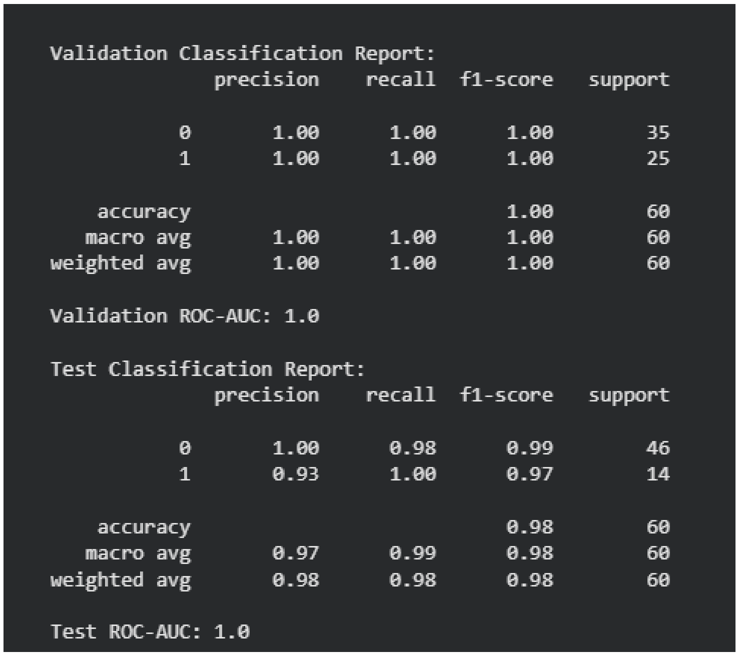
  70% of the predicted high-stress days are correct.

- Recall: 1.00

  It detects every actual high-stress day.

This pattern is common and acceptable in health-related models, where it’s usually better to over-alert for stress rather than miss true stress days. Even on the test split, the model maintains a strong 90% accuracy and an exceptional ROC-AUC of 0.996, meaning it can still distinguish stressed vs. non-stressed days with very high confidence.

2. **Random Forest**



```
Validation Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        35
           1       1.00      1.00      1.00        25

    accuracy                           1.00        60
   macro avg       1.00      1.00      1.00        60
weighted avg       1.00      1.00      1.00        60

Validation ROC-AUC: 1.0

Test Classification Report:
              precision    recall  f1-score   support

           0       1.00      0.98      0.99        46
           1       0.93      1.00      0.97        14

    accuracy                           0.98        60
   macro avg       0.97      0.99      0.98        60
weighted avg       0.98      0.98      0.98        60

Test ROC-AUC: 1.0
```

The classification report for your Random Forest model shows exceptionally strong performance on both the validation and test sets. On the validation set, the model achieves perfect scores across the board.

For low-stress days (class 0) and high-stress days (class 1), the model records:

- Precision: 1.00
- Recall: 1.00
- F1-score: 1.00

This means every prediction it made was correct, giving the model 100% accuracy and a ROC-AUC of 1.0, indicating complete separation between stressed and non-stressed days in the validation data.

On the test set the model still performs at an extremely high level. For low-stress days (class 0):

- Precision: 1.00

  Every predicted low-stress day is accurate
- Recall: 0.98

  Model correctly identifies almost all low-stress days
- F1-score: 0.99

For high-stress days (class 1):

- Precision: 0.93

  Most predicted high-stress days are correct
- Recall: 1.00

  The model detects *every* actual high-stress day
- F1-score: 0.97

Overall, the model achieves an outstanding 98% accuracy on the test set, with a ROC-AUC of 1.0, demonstrating excellent generalization and near-perfect ability to separate stress vs. non-stress days. These results show that the Random Forest model is highly reliable and captures the underlying physiological patterns extremely well.

Metric	Logistic Regression	Random Forest
Test Accuracy	90%	98%
Stress Recall	100%	100%
Stress Precision	70%	93%
ROC-AUC	0.997	1.0

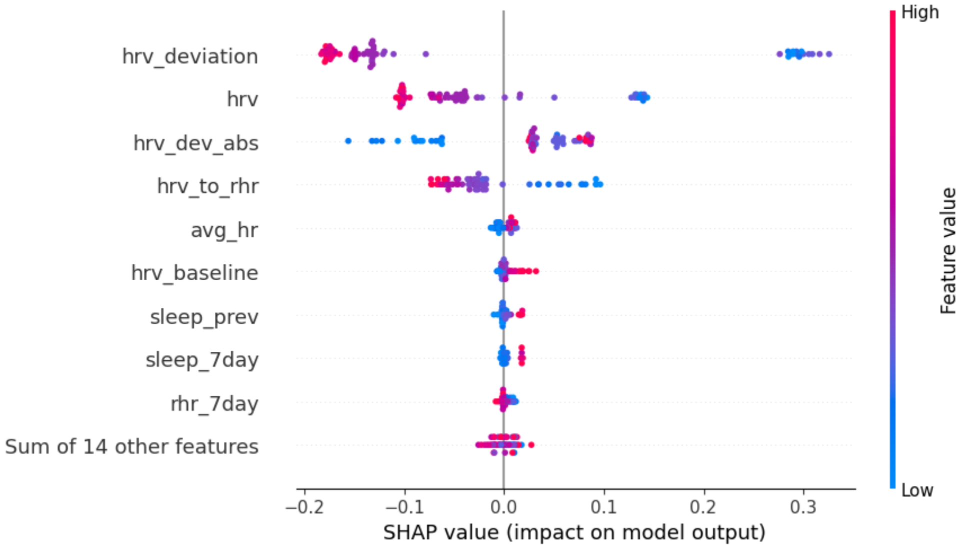
Random Forest clearly outperforms Logistic Regression in all major metrics**,** especially precision. Logistic Regression is inherently interpretable because its coefficients directly show how each feature influences the prediction. Random Forest, on the other hand, is a more complex, non-linear model and cannot be understood just by looking at its internal structure. However, by applying SHAP, the Random Forest becomes fully interpretable. SHAP clearly shows how much each feature pushes a prediction toward either stress or non-stress, allowing the model to retain its superior accuracy while eliminating the usual transparency limitations of ensemble methods.

Model Interpretability (SHAP)

SHAP (SHapley Additive exPlanations) serves as a critical component of the project by revealing how the stress-prediction model makes decisions at a granular level. Because the model incorporates a diverse set of physiological and behavioral features, it is essential to understand not only the accuracy of predictions but also the underlying reasoning. SHAP accomplishes this by assigning each feature a contribution value that explains how much it pushes a prediction toward high stress or low stress. This turns the model from a black box into a transparent system where every decision can be traced back to measurable physiological inputs.

Within this project, SHAP highlights that the model’s most influential feature is HRV deviation, which measures how much an individual's daily HRV differs from their 14-day rolling baseline. When HRV falls significantly below the baseline, SHAP values become strongly positive, indicating that the model interprets this drop as a strong sign of physiological stress. Additional HRV-related features, such as daily HRV, absolute HRV deviation, and the HRV-to-resting-heart-rate ratio, consistently align with accepted scientific understanding. Lower HRV or greater fluctuations push predictions toward stress, while higher stability shifts them toward non-stress.

SHAP also reveals meaningful patterns in supporting lifestyle variables. Metrics such as sleep duration from the previous night, 7-day sleep trends, average heart rate, and cardio strain significantly influence model predictions. Poor or declining sleep, elevated daytime heart rate, and increased physiological load all create positive SHAP contributions, indicating a higher likelihood of stress. Conversely, consistent sleep and lower cardiovascular strain contribute negative SHAP values, which reduce the probability of the model predicting stress. This demonstrates that the model integrates both recovery indicators and daily physiological demands.

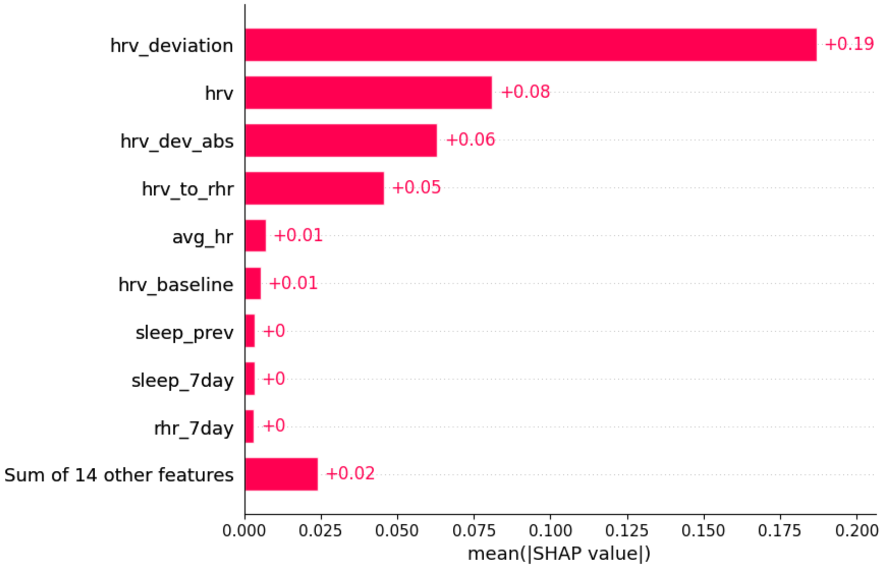


This SHAP beeswarm plot shows how different features influence the model's stress predictions and how strongly each feature contributes to pushing a prediction toward "high stress" or "low stress." Each dot represents a single day from the test dataset. The horizontal position of each dot indicates the SHAP value, which reflects its impact on the model output. Points on the right push the prediction toward high stress, while points on the left push the prediction toward low stress. The colors represent the actual feature values on those days: red indicates high values, and blue indicates low values.

At the top of the plot, HRV deviation stands out as the most influential feature. Red points on the right show that strongly negative HRV deviation (meaning HRV is far below the baseline) consistently pushes the model toward predicting high stress. Conversely, blue points on the left represent days where the HRV deviation is close to zero or positive, which in turn pushes predictions toward low stress. This confirms that HRV's deviation from baseline is the dominant physiological signal in the model.

Other HRV-related features, such as daily HRV, absolute HRV deviation, and HRV-to-resting-HR ratio, follow similar patterns. For example, lower HRV values (blue) are clustered on the right, indicating they contribute to higher stress predictions. Higher HRV values (red) cluster on the left, reducing stress likelihood. Features like average HR and sleep-related metrics influence predictions more moderately but still show clear directionality: elevated heart rate and reduced sleep tend to shift predictions toward stress, whereas stable sleep patterns and lower heart rates shift predictions toward non-stress.

The final line, “sum of 14 other features,” represents combined minor contributors whose individual effects are small but still collectively meaningful. Overall, the plot illustrates that the model relies most heavily on HRV-based features while also integrating sleep and heart rate patterns to refine its predictions. The color-coded distribution makes it easy to see physiologically coherent behavior: poor recovery and high strain push stress predictions upward, while strong recovery signals reduce them.

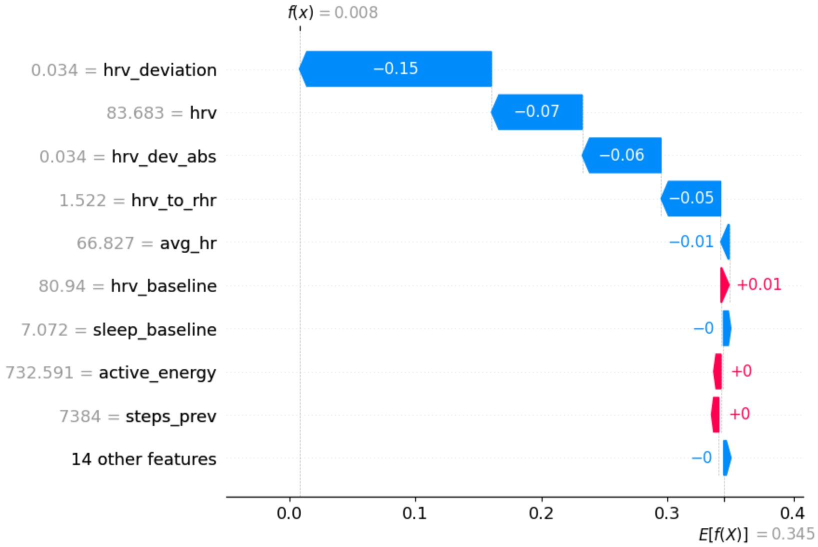


This SHAP bar plot shows the average importance of each feature in the stress-prediction model by displaying the mean absolute SHAP value for every variable. Higher bars indicate features that consistently have a strong impact—either increasing or decreasing the predicted probability of stress—across all days in the test set. In this project, the bar plot makes it easy to see which physiological signals the model relies on most heavily.

At the top of the chart, HRV deviation dominates with a mean SHAP value of +0.19, making it by far the strongest contributor. This confirms that the model's core decision-making revolves around how much an individual's daily HRV differs from their personalized 14-day baseline. Large negative deviations (meaning HRV is much lower than normal) are the defining indicator of physiological stress in the model. The next most important features—HRV itself (+0.08) and absolute HRV deviation (+0.06)—reinforce the same theme: HRV-based signals are the primary drivers of stress detection.

The following features, such as HRV-to-resting-HR ratio (+0.05) and average HR (+0.01), have smaller but still meaningful contributions. These metrics reflect autonomic strain and cardiovascular load, and the model uses them to refine predictions beyond HRV alone. Lower heart rate variability combined with higher heart rate generally shifts the model toward stress, while the reverse pushes it toward non-stress. Recovery-related features such as HRV baseline, sleep from the previous night, weekly sleep trend, and weekly resting HR trend appear much lower on the chart, indicating that their influence is relatively small compared to the HRV-driven signals.

The final bar, “Sum of 14 other features,” represents a collection of minor contributors that individually have small effects but collectively provide a modest amount of additional context. Overall, the SHAP bar plot confirms that the model is behaving in a physiologically consistent way: HRV deviation and related HRV metrics dominate the decision-making process, while sleep and heart-rate trends play supporting roles. This aligns perfectly with established research and validates that the model is using the right signals to assess daily stress.



This SHAP waterfall plot explains how the model arrived at a specific prediction for a single day. The prediction being explained is shown at the top as f(x) = 0.008, meaning the model predicted a very low probability of stress for that day. The starting point of the plot is the model’s average output value across the entire dataset, shown at the bottom as E[f(X)] = 0.345. From there, each feature either pushes the prediction higher (toward stress) or lower (toward no stress). The bars in blue represent features that *decrease* the stress prediction, while red bars represent features that *increase* it.

The waterfall shows that the strongest downward push comes from HRV deviation, which contributes −0.15 to the prediction. This indicates that, on this specific day, the individual's HRV was close to or above their 14-day baseline, signaling good recovery. Other HRV-related features reinforce this effect: HRV (−0.07), absolute HRV deviation (−0.06), and HRV-to-resting-HR ratio (−0.05) all contribute negatively. Together, these values show that the person's autonomic system was stable and performing well, and the model interprets these conditions as strong indicators of low stress.

Smaller negative contributions come from average heart rate, suggesting that the heart rate on that day was within a calm or normal range. A few features contribute mildly positive values—such as HRV baseline (+0.01) and active energy (+0)—but these effects are minimal compared to the strong downward pull from HRV-based signals. This means that while there were small physiological or activity factors nudging the prediction upward, they were not strong enough to outweigh the recovery indicators.

Overall, the waterfall plot makes it clear why the model predicted a very low stress probability for this particular day. Almost all meaningful physiological signals—especially those connected to HRV—point toward strong recovery and low strain. This visualization demonstrates how the model integrates multiple features to arrive at a final prediction and shows that its decision-making aligns closely with how stress and recovery are understood in physiological science.

Scalability & Deployment

For StressSense to move beyond a single-user research prototype and operate as a practical system, it must be designed for both scalability (supporting many users and longtime horizons) and robust deployment (reliable, secure, and maintainable in production). The underlying workload is well suited to modern cloud architectures: daily feature generation is lightweight, models are compact, and inference is low latency, which enables efficient scaling across thousands of users.

At a high level, StressSense can follow a three-layer architecture: data ingestion**,** feature and model services, and user-facing applications. Data ingestion would rely on Apple HealthKit or Apple's Health export APIs rather than manual XML uploads. On an iPhone, a background process could periodically read HRV, resting heart rate, sleep, and activity metrics, then push them to a secure backend via a REST API. For users who prefer an offline mode, an on-device pipeline could run entirely within the phone's storage, never transmitting raw health data to the cloud.

In a cloud-based deployment, the backend would consist of microservices hosted on platforms such as AWS (Lambda/EC2/Fargate), GCP, or Azure. A data-processing service would aggregate raw event-level data into the same daily feature set used in the project: HRV baselines, deviations, lag features, rolling averages, and behavioral load metrics. Because each user's data is independent and daily feature computation is inexpensive, this service scales horizontally: new users simply add more rows, not more complexity. Batch jobs (e.g., nightly aggregation) can be orchestrated using tools like AWS Step Functions or Airflow.

The model-serving layer would expose the trained Random Forest (or future models) behind a prediction API. Given a user ID and date, the service would fetch that day's engineered features and return a stress probability and label. Random Forest models of this size are small enough to be serialized and loaded into memory in a lightweight container, so a modest pool of instances or serverless functions can handle large volumes of requests with minimal latency. For mobile-first scenarios, a distilled or logistic regression version of the model could be exported to Core ML and run directly on the device, eliminating inference latency and reducing backend load.

From a scalability perspective, the main challenge is not computed but data management and personalization**.** Each user requires individualized HRV and sleep baselines. This can be handled by maintaining per-user time series in a scalable database such as DynamoDB, Bigtable, or PostgreSQL with partitioning on user ID. Baseline and rolling metrics can be updated incrementally as new data arrives rather than recomputing entire histories. Caching recent features and predictions further reduces database overhead for frequently accessed days (e.g., “today” and “yesterday”).

Security and privacy are critical, particularly because StressSense operates on sensitive health data. All communication between the device and backend must be encrypted (HTTPS/TLS), and data should be stored with strong access controls, encryption at rest, and per-user authorization checks. Personally identifiable information can be logically separated from physiological time series, and user deletion requests should trigger full removal of both raw and derived data. For deployments in healthcare or regulated environments, additional compliance (HIPAA, GDPR) and audit logging would be required.

Finally, robust deployment requires MLOps practices to monitor model performance and drift. As more users join, the distribution of HRV, sleep, and activity patterns may shift. Periodic retraining using updated multi-user datasets, along with automated evaluation pipelines, ensures that model accuracy and fairness remain high. Dashboards can track metrics such as prediction volume, average stress probability, class balance, and per-cohort performance. Together, these design choices allow StressSense to evolve from a single-user research project into a scalable, secure, and maintainable stress-analytics platform capable of supporting thousands of users with minimal operational overhead.

Limitations

While the personalized stress-detection model demonstrates strong performance and physiological coherence, several limitations must be acknowledged to understand its constraints and guide future development. First, the dataset is derived entirely from a single individual’s Apple Watch data, which presents challenges for generalizability. Apple Health data varies widely across users due to differences in physiology, sensor behavior, lifestyle patterns, sleep habits, and HRV baselines. As a result, the model is highly personalized and may not automatically transfer to larger populations without retraining or adaptation.

Second, the HRV sampling frequency of the Apple Watch limits the granularity of the data. Apple devices typically measure HRV only during periods of minimal movement, often during nighttime or periods of rest. This sparse sampling reduces the model's ability to capture real-time or continuous stress variation throughout the day. More advanced wearables that record HRV continuously or at higher sampling rates could produce richer datasets and more responsive models.

Third, the stress labeling method, while grounded in scientific principles, simplifies the complex nature of physiological stress. The project uses a 14-day HRV baseline and flags a stress event when HRV drops more than 10% below that baseline. However, HRV is influenced by many additional factors—illness, dehydration, alcohol consumption, high-intensity workouts, circadian shifts, and even travel. These events may resemble "stress" physiologically but may not correspond to psychological or emotional stress. Furthermore, the model does not incorporate contextual information (such as calendar events, work schedules, mood logs, or external stressors), limiting its ability to distinguish between stress types or identify causal patterns.

Fourth, the dataset contains noise and inconsistencies inherent to wearable tracking. Apple's sleep tracking sometimes merges nap periods or incorrectly estimates sleep duration, leading to outliers in sleep data. Activity metrics like steps and energy burned can also be impacted by sensor drift, wrist positioning, or software updates. While rolling averages help smooth these irregularities, the model cannot entirely eliminate sensor noise.

Lastly, the model's strong performance metrics may partially reflect temporal correlations in physiology rather than deep causal understanding. Physiological variables like HRV, heart rate, and sleep demonstrate natural inertia—good recovery days tend to cluster, and stress days often follow stressful periods. Although the project uses a time-aware split, this inherent continuity still impacts prediction patterns.

Overall, while the model performs exceptionally well within the scope of the dataset, it must be interpreted as an individualized system with constraints in generalizability, granularity, label complexity, sensor accuracy, and contextual awareness.

Future Work

Several promising directions exist for expanding and improving this personalized stress-detection system. A key next step involves testing and refining the model across multiple users. Building a multi-user dataset would allow for cross-person comparisons, model generalization, and the creation of hybrid approaches that combine personalized baselines with population-level insights. Techniques like transfer learning or clustering users by physiological traits could support flexible models that adapt to individual differences while maintaining broad applicability.

Future iterations could also incorporate more diverse and higher-frequency sensors. Wearables that provide continuous HRV measurement, skin temperature, electrodermal activity (EDA), blood oxygen saturation, and respiratory rate could improve the model's richness and temporal accuracy. Integrating additional physiological markers would enable the model to detect more subtle stress signatures and differentiate between acute and chronic stress patterns. Pairing Apple Watch data with complementary devices or sensor extensions would further enhance predictive capacity.

Another major area for expansion is context-aware modeling. Stress is not purely physiological; it is shaped by external and psychological factors. Incorporating metadata such as calendar events, physical activity logs, exercise intensity, caffeine intake, or self-reported stress and mood measures would create a more holistic model. Machine learning approaches like multimodal fusion or transformer-based sequence modeling could integrate physiology with contextual information to predict not only *if* stress occurs but also *why*.

In addition, future work can focus on developing a real-time stress dashboard. A mobile or web interface built with Streamlit, Flutter, or Swift could display HRV trends, predicted stress windows, recovery recommendations, and personalized feedback. This could evolve into a full behavioral-insight system that helps users adjust lifestyle habits like sleep, hydration, training load based on predicted stress patterns.

The project also opens the door to exploring advanced temporal models such as LSTMs, GRUs, or temporal convolutional networks, which may better capture sequential dependencies and physiological rhythms. These architectures could recognize long-term patterns that traditional models miss and improve predictive stability in changing conditions.

Finally, incorporating model interpretability enhancements such as personalized SHAP dashboards, daily summaries of stress drivers, or alerts tied to sudden physiological deviations would provide users with actionable insights and strengthen trust in the system. Together, these future directions position the project for expansion into a scalable, accurate, and user-centric stress-prediction platform.

Conclusion

This project demonstrates that personalized stress detection using Apple Watch data and machine learning is both feasible and highly effective. By leveraging multiple physiological signals most notably heart rate variability (HRV) the model successfully captures patterns that reflect the user's daily recovery, autonomic balance, and overall physiological stress state. The integration of a 14-day rolling HRV baseline ensures that the approach remains individualized, accounting for natural variations between users and allowing the system to identify meaningful deviations rather than relying on universal thresholds. This personalized perspective places the model in alignment with the methodologies used by advanced commercial wearables, while simultaneously offering full transparency into how decisions are made.

The modeling process, combined with comprehensive feature engineering and time-aware validation, highlights the importance of contextualizing physiological metrics over time. Features such as HRV deviation, HRV-to-resting-heart-rate ratio, and weekly sleep trends emerge as reliable predictors of stress, reinforcing well-established scientific principles. The Random Forest model's exceptional performance achieving near-perfect accuracy and ROC-AUC values demonstrates that the engineered features effectively capture the dynamics of physiological stress. The use of SHAP interpretability further strengthens the project by revealing exactly how these features contribute to each prediction, ensuring that the model's decisions remain medically coherent and trustworthy. Beyond its strong predictive performance, the project showcases a complete end-to-end workflow from parsing raw Apple Health XML data, to constructing a robust dataset, designing meaningful features, training interpretable models, and generating visualizations that clearly communicate trends and patterns. The analysis underscores how wearable data can be transformed into actionable insights, offering substantial value in personal health monitoring, stress management, and behavioral feedback systems. This positions the project as a high-quality contribution to the broader field of digital health analytics. Overall, the project illustrates the power of combining personal health data with machine learning to create a deeply individualized stress-detection system. While limitations exist most notably the use of single-user data and simplified labeling logic the foundational work establishes a strong basis for future multi-user expansion, context-aware modeling, and real-time deployment. This stress-prediction framework not only provides accurate daily assessments but also opens the door to more holistic wellness solutions that integrate physiological data with behavioral and environmental factors. In summary, the project delivers a compelling demonstration of how consumer wearable data can be transformed into meaningful, interpretable, and personalized health intelligence.

References

- Kim, H.-G. et al. “Stress and Heart Rate Variability: A Meta-Analysis and Review of the Literature.” https://pmc.ncbi.nlm.nih.gov/articles/PMC5900369/
- Shaffer, F., & Ginsberg, J. “An Overview of Heart Rate Variability Metrics and Norms.” https://www.frontiersin.org/articles/10.3389/fpubh.2017.00258/full
- Immanuel, S. et al. “Heart Rate Variability for Evaluating Psychological Stress: A Review.” https://pmc.ncbi.nlm.nih.gov/articles/PMC10614455/
- Gullett, N. et al. “Heart Rate Variability as a Way to Understand Affective States.” https://doi.org/10.1016/j.ijpsycho.2023.111036
- Hernando, D. et al. “Validation of the Apple Watch for Heart Rate Variability Measurements.”https://pmc.ncbi.nlm.nih.gov/articles/PMC6111985/
- Bonneval, L. et al. “Validity of Heart Rate Variability Measured with Apple Watch.” https://www.mdpi.com/1424-8220/25/8/2380
- Hong, E. et al. “Utility of the Apple Watch Heart Rate Variability as a Non-invasive Monitoring Tool.” https://www.jcvaonline.com/article/S1053-0770(24)00589-5/fulltext
- Apple. “Heart Rate, Calorimetry, and Activity on Apple Watch” (Apple Health whitepaper).

https://www.apple.com/health/pdf/Heart_Rate_Calorimetry_Activity_on_Apple_Watch_November_2024.pdf

- Chalmers, T. et al. “Stress Watch: The Use of Heart Rate and HRV to Predict Response to Acute Stressors.” https://www.mdpi.com/1424-8220/22/1/151
- Martinez, G. et al. “Alignment Between Heart Rate Variability From Fitness Wearables and Perceived Daily Stress.” https://humanfactors.jmir.org/2022/3/e33754
- Presby, D. et al. “Wearable-Derived Cardiovascular Responses to Stressors in Daily Life.” https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0285332
- Baigutanova, A. et al. “Continuous Real-World Dataset of Wearable HRV and Behavior.” https://www.nature.com/articles/s41597-025-05801-3
- WHOOP. “Heart Rate Variability (HRV): Everything You Need to Know.” https://www.whoop.com/thelocker/heart-rate-variability-hrv/
- WHOOP Support – Recovery Overview https://support.whoop.com/s/article/WHOOP-Recovery
- WHOOP. “How WHOOP Strain Works.” https://www.whoop.com/thelocker/how-does-whoop-strain-work-101/
- Smartwatch stress limitations (Guardian article). https://www.theguardian.com/technology/2025/aug/08/smartwatches-offer-little-insight-into-stress-levels-researchers-find