Ensemble Learning for Stress Prediction Using Wearable devices

# **Introduction**

Stress, often referred to as a silent killer, poses a significant threat to human health and well-being. Its insidious nature has been linked to the exacerbation of serious illnesses, such as diabetes, heart disease, and high blood pressure. Startling statistics from the British Health and Safety Executive in 2021-2022 revealed that stress was responsible for a staggering 50% of all work-related diseases, highlighting the urgency of addressing this modern epidemic.

The detrimental effects of stress on both physical and mental health have been extensively studied and well-documented. Epel et al. (2018) conducted research that underscored the profound impact of stress on individuals' overall well-being. While young and resilient individuals with adaptive coping mechanisms might be better equipped to handle short-term stress, prolonged or intense stress experiences increase the likelihood of developing chronic disorders, often associated with depression and anxiety (Arza et al., 2019). It is essential to recognize that stress is not limited to causing acute events, such as heart attacks or strokes, but also plays a significant role in chronic conditions. Scientific investigations led by Tawakol et al. (2017) have established a direct link between chronic stress and life-threatening diseases like heart disease, high blood pressure, diabetes, and obesity. The accumulating evidence on stress-related health issues calls for proactive measures to address and mitigate its impact on society.

Traditionally, in clinical settings, self-reported questionnaires, like the Perceived Stress Scale (PSS), have been employed to assess the subjective experience of stress in individuals (Reis et al., 2010). While these questionnaires have proven to be valuable tools, recent technological advancements have introduced less invasive and more accurate methods for monitoring stress levels – wearable devices. Wearable devices are a revolutionary innovation equipped with a diverse array of sensors, including temperature sensors, accelerometers, optical sensors, and biometric sensors, that enable the continuous monitoring of various physiological signals. Although some of these sensors may not currently match the precision of fixed hospital equipment, they are considered acceptable for relevant applications (Hayano et al., 2020; Delmastro et al., 2020). The incorporation of wearable devices into stress monitoring not only offers a less intrusive method but also holds the potential for a more comprehensive and dynamic assessment of an individual's stress levels. Such technology empowers individuals to gain insights into their stress patterns, helping them adopt effective coping strategies and make informed lifestyle changes.

One of the significant advantages of wearable devices is their ability to provide real-time data, which allows individuals to identify stressful triggers and promptly intervene. Understanding how stress manifests in different situations enables people to develop personalized stress management techniques, fostering a healthier response to life's challenges. Furthermore, the continuous monitoring of stress levels through wearables can provide valuable data for healthcare professionals and researchers. By aggregating anonymized stress data from diverse populations, scientists can gain a deeper understanding of stress-related trends, risk factors, and potential interventions. This collective knowledge can be instrumental in designing targeted public health campaigns and policies to address the widespread impact of stress on society.

As wearable technology continues to evolve, we can expect further improvements in the precision and sophistication of stress monitoring capabilities. The integration of artificial intelligence and machine learning algorithms may enable wearables to not only measure stress levels accurately but also predict stress episodes based on individual patterns and contexts. However, the growing adoption of wearable devices for stress monitoring also raises concerns about data privacy and security. As these devices gather sensitive health information, it becomes imperative to implement robust safeguards to protect users' personal data from unauthorized access or misuse. Striking a balance between leveraging the benefits of wearable technology and safeguarding individual privacy is a challenge that needs careful consideration.

Another promising area of research in stress monitoring lies in the potential synergy between wearables and modern artificial intelligence; this presents a promising frontier in stress monitoring and prediction. As AI technologies continue to advance, they have the potential to revolutionize the way we manage stress and enhance overall well-being. By combining wearable devices with AI-powered algorithms, individuals can benefit from more accurate and personalized stress management programs. Wearable devices equipped with various sensors can continuously monitor physiological signals, providing a wealth of data on an individual's stress levels and responses to different situations. AI algorithms can then analyze this data, identifying patterns and correlations that may not be immediately evident to the human eye.

The integration of wearable devices and AI in telemedicine applications is another avenue that holds great potential. Telemedicine allows healthcare professionals to remotely monitor patients, providing personalized care and timely interventions. By incorporating wearable devices into telemedicine platforms, healthcare providers can gain insights into patients' stress levels and responses, even from a distance. This remote monitoring capability is particularly beneficial for individuals with chronic stress-related conditions who require ongoing support and management. Through AI-enabled analysis of wearable data, healthcare professionals can tailor treatment plans and interventions, optimizing patient outcomes and quality of life.

Apologies for the confusion. Let's revise the write-up to accurately reflect the stress prediction project using the SWELL dataset and synthetic data generated through numerical simulation:

This dissertation focuses on a stress prediction project utilizing two distinct datasets to develop and evaluate predictive models. The primary dataset employed is the SWELL dataset, consisting of heart rate variability (HRV) indices computed from the multimodal SWELL knowledge work (SWELL-KW) dataset. The SWELL-KW dataset was collected by researchers at the Institute for Computing and Information Sciences at Radboud University. It involved experiments conducted on 25 subjects engaged in typical office work, experiencing various stress-inducing situations, such as receiving unexpected emails, interruptions, and time pressure to complete tasks. The dataset contains a rich array of data, including computer logging, facial expressions, body postures, ECG signals, and skin conductance. In addition to the SWELL dataset, this dissertation leverages synthetic data generated through numerical simulation. The synthetic data is designed to resemble real-world stress-related scenarios and is used to augment the existing dataset. The use of synthetic data allows for the exploration of a broader range of stress conditions and ensures a comprehensive evaluation of stress prediction models.

The document is organized in the following sections: (i) a literature review where related research is presented, (ii) a description of the data and the preprocessing tasks done, (iii) a description of the methodology (iv) a presentation of the results obtained (v) some conclusions and proposals for future work.

# **Literature Review**

Stress is a pervasive and concerning issue that significantly impacts individuals' health and well-being. As research on stress continues to evolve, one particularly promising area gaining traction is stress prediction using data collected from wearable devices. These innovative devices, equipped with a variety of sensors, enable continuous monitoring of physiological signals, providing real-time data on an individual's stress levels. The integration of wearable technology with stress prediction algorithms powered by artificial intelligence (AI) holds great potential for personalized and proactive stress management strategies.

**2.1 Stress Assessment Methods**

This section will review traditional stress assessment methods, such as self-reported questionnaires, and discuss their limitations in capturing real-time stress responses.

Traditional stress assessment methods have been widely employed, such as self-reported questionnaires and interviews, to capture individuals' perceived stress levels. These methods rely on participants' ability to introspect and report their emotional and cognitive experiences accurately. One commonly used questionnaire is the Perceived Stress Scale (PSS) developed by Cohen et al. (1983), which measures the degree to which situations in life are appraised as stressful. While self-reported measures have been valuable in understanding subjective experiences of stress, they have several limitations that hinder their ability to capture real-time stress responses. One significant limitation of self-reported questionnaires is their reliance on retrospective accounts of stress, which may lead to recall biases and inaccuracies. Human memory is fallible, and stress experiences can be complex and dynamic, making it challenging for individuals to accurately recall their stress levels over time. Moreover, self-report measures are susceptible to social desirability biases, where participants may provide responses they perceive as more socially acceptable, leading to inaccuracies in the data.

Another challenge with traditional stress assessment methods is their inability to provide real-time monitoring of stress responses. Stress is a highly dynamic process that fluctuates throughout the day, influenced by various situational and environmental factors. Traditional methods lack the sensitivity to capture these moment-to-moment changes in stress, making it difficult to gain a comprehensive understanding of individuals' stress patterns and triggers. To address these limitations and advance stress assessment, researchers have turned to wearable devices and physiological monitoring. Wearable technology, such as heart rate monitors, electrodermal activity sensors, and accelerometers, offers a non-invasive and unobtrusive way to continuously track physiological signals associated with stress responses. These devices can provide real-time data on heart rate variability, skin conductance, body movement, and other indicators that change in response to stress.

**2.2 Wearable Devices in Stress Monitoring**

In this section, the focus will be on examining the technological advancements in wearable devices and the sensors they employ for stress monitoring.

Numerous studies have explored the feasibility of using wearable devices for stress assessment. For example, Gjoreski et al. (2016) employed a wrist-worn sensor to measure heart rate and skin conductance in combination with a smartphone app to detect stress levels during daily activities. Their findings demonstrated that the combination of physiological signals and machine learning algorithms could achieve promising accuracy in distinguishing stressful situations from non-stressful ones. AI algorithms play a pivotal role in harnessing the potential of wearable devices for stress prediction. Machine learning techniques, such as support vector machines, neural networks, and random forests, have been used to process the vast amounts of data generated by wearables and extract meaningful patterns associated with stress. These algorithms can learn from labeled datasets and develop models capable of predicting stress episodes based on physiological signals.

For instance, a study by LiKamWa et al. (2018) utilized deep learning techniques to analyze physiological data collected from wearable sensors and predict stress levels with high accuracy. The authors demonstrated that their model could effectively identify stress patterns, offering insights into personalized stress management strategies. Moreover, advancements in AI and wearable technology have enabled the development of real-time stress prediction systems. These systems can continuously monitor physiological signals, process the data on the device or through cloud computing, and provide timely feedback to users when stress levels exceed certain thresholds. Such applications hold great potential in promoting stress-awareness and proactive stress management.

**2.3 AI-Powered Stress Prediction**

This section will explore the integration of AI algorithms in stress prediction models using wearable data.

The integration of AI algorithms in stress prediction models using wearable data represents a groundbreaking approach to revolutionizing stress assessment and management. By leveraging the power of machine learning and deep learning techniques, researchers and developers aim to harness the vast amounts of physiological data collected from wearable devices to create accurate and personalized stress prediction models. One of the key challenges in stress prediction is the complexity and variability of physiological signals associated with stress responses. Wearable devices continuously capture data such as heart rate, skin conductance, body movement, and sleep patterns, which can be influenced by a range of factors beyond stress, including physical activity, environmental conditions, and individual differences. Consequently, developing effective prediction models requires sophisticated AI algorithms capable of discerning stress-related patterns amidst this noise and variability.

Machine learning techniques have emerged as valuable tools in stress prediction models. Supervised learning, in particular, involves training algorithms on labeled datasets, where physiological data is paired with corresponding stress levels or stress episodes. These algorithms learn from this data to identify patterns and relationships between physiological signals and stress states. Support vector machines (SVM), decision trees, and random forests are some of the classical supervised learning algorithms that have been applied in stress prediction studies.

For instance, a study by Jain et al. (2020) utilized an SVM-based approach to predict stress episodes based on heart rate variability and skin conductance data collected from wearable sensors. The model achieved high accuracy in distinguishing stressful events from non-stressful ones, providing a promising step towards real-time stress prediction. In recent years, deep learning algorithms, specifically neural networks, have gained traction in stress prediction due to their ability to handle complex and high-dimensional data. Deep learning models can automatically learn hierarchical representations of features from raw physiological data, potentially capturing subtle patterns indicative of stress that may be challenging for traditional machine learning techniques to detect.

A notable example is the work of Chen et al. (2022), who employed a recurrent neural network (RNN) to analyze continuous heart rate and accelerometer data collected from smartwatches. Their model not only predicted stress episodes but also identified patterns of stress development over time, allowing for more comprehensive stress assessment and management. Furthermore, the integration of AI algorithms in stress prediction models enables personalized stress monitoring and intervention strategies. AI-driven models can learn individual differences in stress responses and tailor predictions based on each user's unique physiological profile. This personalized approach can provide users with targeted feedback and coping strategies, enhancing the effectiveness of stress management interventions.

While AI algorithms hold immense promise in stress prediction using wearable data, it is essential to address several challenges and ethical considerations. Firstly, the size and quality of the training data significantly influence the performance of AI models. Large and diverse datasets are required to train robust and generalizable prediction models. Collaboration between researchers, data scientists, and wearable device manufacturers is crucial in creating comprehensive datasets for stress prediction research. Secondly, transparency and interpretability of AI models are critical, especially in the healthcare domain. Understanding how the models arrive at their predictions is essential for building trust and facilitating the adoption of AI-driven stress prediction systems by both users and healthcare professionals.

**2.4 Applications and Implications**

In this section, the focus will be on discussing the practical applications and implications of stress prediction using wearable devices.

Stress prediction using wearable devices and AI algorithms has far-reaching applications and significant implications across various domains. This innovative approach to stress assessment opens up new possibilities for personalized interventions, healthcare advancements, workplace well-being, and overall societal impact.

1. Personalized Stress Management:

One of the most immediate applications of stress prediction using wearables is in personalized stress management. By continuously monitoring physiological signals, wearable devices can provide real-time feedback to individuals about their stress levels. This information empowers users to identify stress triggers, patterns, and potential stressors in their daily lives. Armed with this knowledge, individuals can adopt targeted coping strategies, such as mindfulness exercises, deep breathing techniques, or physical activity, to mitigate stress and promote well-being (LiKamWa et al., 2018).

1. Healthcare and Mental Health Interventions:

The integration of AI-driven stress prediction models into healthcare settings holds great potential for improving mental health interventions. Healthcare providers can use wearable devices to monitor stress levels in patients with conditions like anxiety disorders, depression, or post-traumatic stress disorder (PTSD). Real-time stress data can aid in tailoring treatment plans and tracking the effectiveness of therapeutic interventions over time. Early detection of heightened stress levels may also enable timely interventions to prevent the escalation of mental health issues (Chen et al., 2022).

1. Workplace Stress Management:

Work-related stress is a prevalent issue affecting employee well-being and productivity. Wearable devices with stress prediction capabilities can be utilized in the workplace to monitor employees' stress levels. Employers can use this data to identify high-stress periods and implement targeted stress reduction programs or adjust workloads accordingly. Moreover, employees themselves can use wearable devices to better manage their stress and maintain work-life balance (Gjoreski et al., 2016).

1. Public Health and Well-Being:

Aggregated data from wearable devices can provide valuable insights into population-level stress trends and patterns. Public health officials can use this information to understand stress-related issues affecting communities and design targeted interventions to promote overall well-being. Additionally, stress prediction models can be integrated into health promotion campaigns, encouraging individuals to proactively manage their stress and lead healthier lives (Jain et al., 2020).

1. Early Warning Systems:

AI-driven stress prediction models can serve as early warning systems for individuals at risk of experiencing chronic stress or stress-related health issues. By identifying subtle physiological changes that may precede more severe stress episodes, wearable devices can prompt individuals to seek timely medical advice or implement preventive measures, potentially reducing the risk of stress-related health complications (LiKamWa et al., 2018).

1. Ethical and Social Implications:

The widespread adoption of stress prediction using wearables and AI also raises ethical and social considerations. Privacy and data security must remain paramount, as sensitive physiological data is continuously collected from individuals. Transparent communication about data usage, informed consent, and data anonymization are essential to ensure users' trust and safeguard their privacy (Chen et al., 2022).

1. Health Disparities and Access:

The integration of wearable devices and AI in stress prediction has the potential to exacerbate health disparities if access to such technologies is unequal. Efforts should be made to ensure that individuals from diverse socioeconomic backgrounds have access to these tools, preventing further marginalization and inequity in stress management and mental health care (Gjoreski et al., 2016).

# **Data**

In this project, we focus on the data utilized, which comprises several key aspects. Firstly, we provide a concise overview of the SWELL-KW dataset. Secondly, a comprehensive description of the SWELL-KW dataset is presented, covering its collection process, as well as details about the synthetic data generation, including the procedure, assumptions, and criteria used during its creation. Thirdly, we delve into the data cleaning and preprocessing tasks carried out, along with the analysis performed. Lastly, we showcase the aggregated dataset, which we employ for the stress prediction process.

## **SWELL-KW Dataset**

The SWELL-KW dataset serves as a crucial resource for stress prediction in this research on user modeling. It comprises heart rate variability (HRV) indices computed from the multimodal data collected during experiments conducted at the Institute for Computing and Information Sciences at Radboud University.

The dataset was gathered by involving 25 subjects in typical office work, such as report writing, presentations, email reading, and information searching. Throughout the experiments, the participants encountered various stress-inducing scenarios, including unexpected email interruptions and time pressure to complete tasks. The recorded data encompassed diverse aspects, including computer logging, facial expressions, body postures, ECG signals, and skin conductance. Additionally, subjective experiences related to task load, mental effort, emotions, and perceived stress were recorded.

During the experiments, each participant experienced three distinct working conditions:

1. No stress: Subjects had unrestricted time to complete tasks, unaware of any maximum time limit.
2. Time pressure: The time to finish tasks was reduced to 2/3 of the time taken in the no-stress condition.
3. Interruption: Participants received eight emails during their tasks, some relevant and requiring specific actions, while others were irrelevant.

To compute the HRV indices, the researchers adopted a meticulous approach. They extracted the inter-beat interval (IBI) signal from the peaks of the subjects' electrocardiography (ECG) data. The HRV indices were then computed based on 5-minute IBI arrays. This process allowed for a more detailed analysis of how each heartbeat influenced the person's HRV, proving superior to traditional methods that analyzed HRV on the whole signal.

The research contribution of this dataset lies in its comprehensiveness compared to other datasets. By using this enriched dataset, stressors can be predicted with remarkable accuracy, reaching an impressive 99.25%.

Below are the columns in this dataset and their description

| **Feature Name** | **Description** | **Formula** |
| --- | --- | --- |
| MEAN\_RR | Mean of all RR intervals | Mean(RR\_intervals) |
| MEDIAN\_RR | Median of all RR intervals | Median(RR\_intervals) |
| SDRR | Standard deviation of all intervals | StandardDeviation(RR\_intervals) |
| RMSSD | Square root of the mean of the sum of the squares | SquareRoot(Mean((RR\_intervals - Mean(RR\_intervals))^2)) |
| SDSD | Standard deviation of all interval differences | StandardDeviation(Diff(RR\_intervals)) |
| SDRR\_RMSSD | Ratio of SDRR over RMSSD | SDRR / RMSSD |
| HR | Heart Rate (beats per minute) | 60 / MEAN\_RR |
| pNN25 | Percentage of adjacent RR intervals differing by more than 25 ms | (Number of intervals differing > 25 ms) / Total intervals |
| pNN50 | Percentage of adjacent RR intervals differing by more than 50 ms | (Number of intervals differing > 50 ms) / Total intervals |
| SD1 | Poincaré plot descriptor of short-term HRV | StandardDeviation(SD1) |
| SD2 | Poincaré plot descriptor of long-term HRV | StandardDeviation(SD2) |
| KURT | Kurtosis of all RR intervals | Kurtosis(RR\_intervals) |
| SKEW | Skewness of all RR intervals | Skewness(RR\_intervals) |
| MEAN\_REL\_RR | Mean of all relative RR intervals | Mean(relative\_RR\_intervals) |
| MEDIAN\_REL\_RR | Median of all relative RR intervals | Median(relative\_RR\_intervals) |
| SDRR\_REL\_RR | Standard deviation of all relative RR intervals | StandardDeviation(relative\_RR\_intervals) |
| RMSSD\_REL\_RR | Square root of the mean of the sum of the squares | SquareRoot(Mean((relative\_RR\_intervals - Mean(relative\_RR\_intervals))^2)) |
| SDSD\_REL\_RR | Standard deviation of all interval differences | StandardDeviation(Diff(relative\_RR\_intervals)) |
| SDRR\_RMSSD\_REL | Ratio of SDRR\_REL over RMSSD\_REL | SDRR\_REL / RMSSD\_REL |
| KURT\_REL\_RR | Kurtosis of all relative RR intervals | Kurtosis(relative\_RR\_intervals) |
| SKEW\_REL\_RR | Skewness of all relative RR intervals | Skewness(relative\_RR\_intervals) |
| VLF | Very low (0.003Hz - 0.04Hz) frequency band of HRV power spectrum | Power in VLF frequency band |
| LF | Low (0.04Hz - 0.15Hz) frequency band of HRV power spectrum | Power in LF frequency band |
| HF | High (0.15Hz - 0.4Hz) frequency band of HRV power spectrum | Power in HF frequency band |
| TP | Total HRV power spectrum | Power in VLF + Power in LF + Power in HF |
| LF/HF | Ratio of LF to HF | LF / HF |
| HF/LF | Ratio of HF to LF | HF / LF |
| sampen | Sample entropy of the RR signal | Calculate Sample Entropy using RR signal |
| higuci | Higuchi Fractal Dimension | Calculate Higuchi Fractal Dimension using RR signal |

## **Wearable Device Simulated Data**

For the second dataset, we opted for a simulated recording emulating typical data collected from an Apple Watch. This simulated dataset comprises 10 different features extracted from the wearable device, each carefully generated to reflect real-world characteristics. Let's delve into the details of these features:

1. Heart Rate: This feature represents continuous monitoring of heart rate throughout the day. The unit of measurement is beats per minute (bpm). In the generated dataset, heart rate values were simulated using a normal distribution with a mean and standard deviation that reflect typical resting heart rate values. This assumption is based on the understanding that heart rate can vary based on factors like physical activity, stress levels, and overall health.
2. Heart Rate Variability (HRV): HRV measures the variation in time intervals between successive heartbeats. It provides insights into the autonomic nervous system's activity and can be an indicator of stress and overall well-being. The unit of measurement is milliseconds (ms). In the generated dataset, HRV values were simulated using a log-normal distribution, which is commonly observed in HRV data. The mean and standard deviation of the log-normal distribution were adjusted to reflect typical HRV values.
3. Respiratory Rate: This feature measures the number of breaths per minute. The unit of measurement is breaths per minute (bpm). In the generated dataset, respiratory rate values were simulated using a normal distribution with mean and standard deviation values that represent the typical range of respiratory rates. It is important to note that respiratory rate can vary depending on factors such as physical exertion, respiratory health, and emotional state.
4. Skin Conductance: Skin conductance measures the electrical conductance of the skin, which can indicate stress levels and emotional arousal. The unit of measurement is microsiemens (µS). In the generated dataset, skin conductance values were simulated using a gamma to capture the variability often observed in skin conductance data. The distribution parameters were adjusted to approximate real-world values.
5. Body Movement: This feature captures body movement using accelerometer data. The generated dataset includes three separate columns (X, Y, and Z) representing movement along different axes. The unit of measurement is typically in g-force or gravitational units. In the generated dataset, body movement values were simulated using random noise and periodic patterns to mimic real-world movement patterns.
6. Sleep Duration: Sleep duration represents the total duration of sleep. The unit of measurement is hours. In the generated dataset, sleep duration values were simulated using a normal distribution with mean and standard deviation values that represent the typical range of sleep durations. This assumption is based on the understanding that sleep duration can vary among individuals and is influenced by factors such as age, lifestyle, and sleep quality.
7. Sleep Efficiency: Sleep efficiency measures the percentage of time spent asleep compared to the total time spent in bed. It is a measure of sleep quality. In the generated dataset, sleep efficiency values were simulated using a normal distribution. This distribution captures the range of possible values from 0% to 100%. Sleep efficiency can be influenced by various factors, including sleep disorders, environmental conditions, and personal habits.
8. Active Energy Burned: This feature tracks the calories burned through physical activity. The unit of measurement is calories (cal). In the generated dataset, active energy burned values were simulated using a normal distribution with mean and standard deviation values that reflect typical energy expenditure patterns for various activities. The assumption is that energy burned varies based on activity intensity, duration, and individual characteristics.
9. Stand Hours: Stand hours represent the number of hours a person spends standing. In the generated dataset, stand hours values were simulated using a bimodal distribution. This distribution reflects the typical distribution of standing time throughout the day, with two peaks representing periods of standing during working hours and leisure time.
10. Ambient Noise Level: Ambient noise level measures the surrounding noise level throughout the day. The unit of measurement is usually in decibels (dB). In the generated dataset, ambient noise level values were simulated using a skewed distribution, a gamma distribution. This choice of distribution accounts for the common observation that ambient noise levels can vary throughout the day and often exhibit skewed characteristics.

By using more realistic distributions and assumptions for each feature, the generated dataset better approximates the characteristics of the real-world data. These adjustments enable a more accurate representation of the features' distribution, variability, and relationships, enhancing the predictive power of the ensemble model.

**3.2.1 Stress Score Calculation**

The process of creating the stress score involved several steps, including assigning weights to the features, normalizing the data, and calculating the stress score. Here is a detailed write-up of the process:

**Assigning Weights to Features**

Each feature was assigned a weight that reflected its relative importance in determining stress levels. The weights were determined based on general accepted standards and knowledge about the importance of each feature. These weights were assigned as follows:

* Heart Rate: 0.15
* HRV: 0.1
* Respiratory Rate: 0.08
* Skin Conductance: 0.12
* Body Movement: 0.07
* Sleep Duration: 0.1
* Sleep Efficiency: 0.05
* Active Energy Burned: 0.18
* Stand Hours: 0.1
* Ambient Noise Level: 0.05

**Data Normalization**

Before calculating the stress score, it was important to normalize the data to ensure that each feature contributed proportionally to the final score. The normalization process involved scaling the values of each feature to a common range (typically between 0 and 1) using the formula:

normalized\_value = (original\_value - min\_value) / (max\_value - min\_value)

This normalization step ensures that the features are on a comparable scale and avoids any bias that may arise from differences in the magnitude of the feature values.

**Calculating the Stress Score**

Once the data was normalized, the stress score was calculated using a weighted sum approach. The stress score for each data point was obtained by taking the dot product between the normalized feature values and their corresponding weights.

stressScore = np.dot(normalizedData.values, np.array(list(weights.values())))

**Incorporating the Stress Score**

Finally, the calculated stress score was added as a new column, 'stress', in the DataFrame to provide a quantifiable measure of stress for each data point. The stress score provides valuable insights into the overall stress levels associated with different combinations of feature values and serves as a useful metric for predicting stress levels based on the features.

## **Preprocessing of the Dataset**

The only preprocessing done was on the SWELL-KW dataset and only involves converting the condition row which was in a non numeric format before.

## **Exploratory Data Analysis**

**SWELL-KW Dataset**

The dataset includes three distinct conditions under which the subjects performed their tasks: "No Stress," "Time Pressure," and "Interruption." Each condition is explained below:

1. No Stress: In the "No Stress" condition, the subjects were given the freedom to work on their tasks without any time constraints. They could take as much time as they needed to complete the tasks, with one notable limitation—a maximum duration of 45 minutes for each task. However, the participants were not aware of this time limit, allowing them to work without feeling pressured by time constraints. This condition aimed to observe the subjects' natural workflow and performance without the influence of external stressors.
2. Time Pressure: Under the "Time Pressure" condition, the participants faced a challenging scenario where the time to complete the tasks was significantly reduced. Specifically, the time given to finish the tasks was set to two-thirds of the time the participant took in the "No Stress" condition for the same tasks. This manipulation aimed to induce a sense of urgency and time-related stress, pushing the subjects to work faster and efficiently.
3. Interruption: The "Interruption" condition involved introducing external disruptions while the subjects were engaged in their assigned tasks. Throughout this condition, participants received a total of eight emails in the middle of their work. Some of these emails were directly related to their ongoing tasks, requiring specific actions from the participants. On the other hand, some emails were entirely unrelated and had no connection to their work. This condition aimed to assess how the subjects coped with task interruptions, multitasking demands, and the impact of shifting focus on their stress levels.

The primary target variable to be predicted in this study is the "Condition" column, which corresponds to the three conditions— "No Stress," "Time Pressure," and "Interruption."

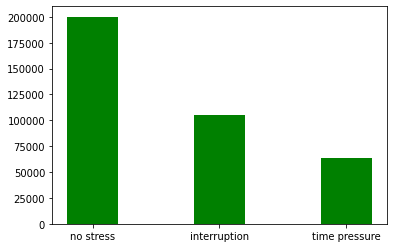
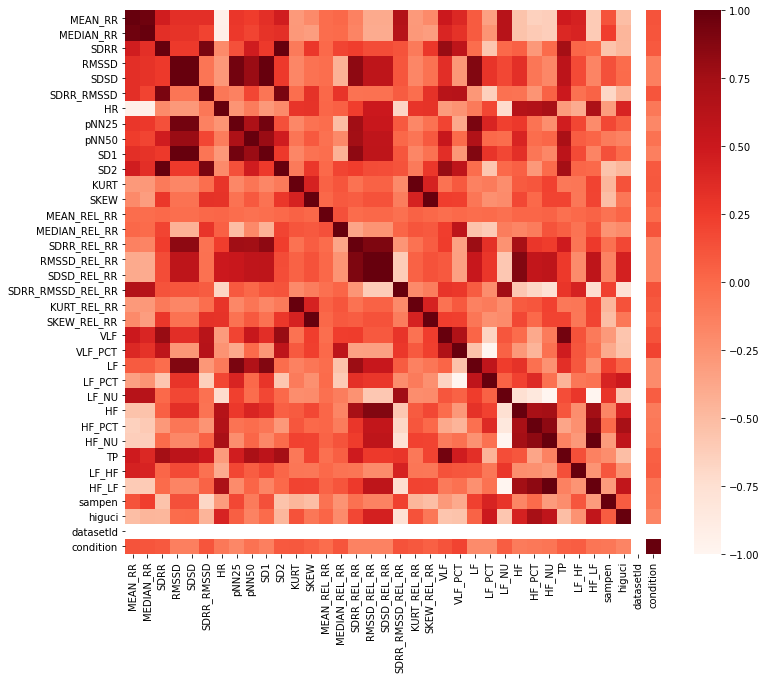
Figure 1: Visualization of the count of each condition

Figure 2: Correlation of all columns in the dataset

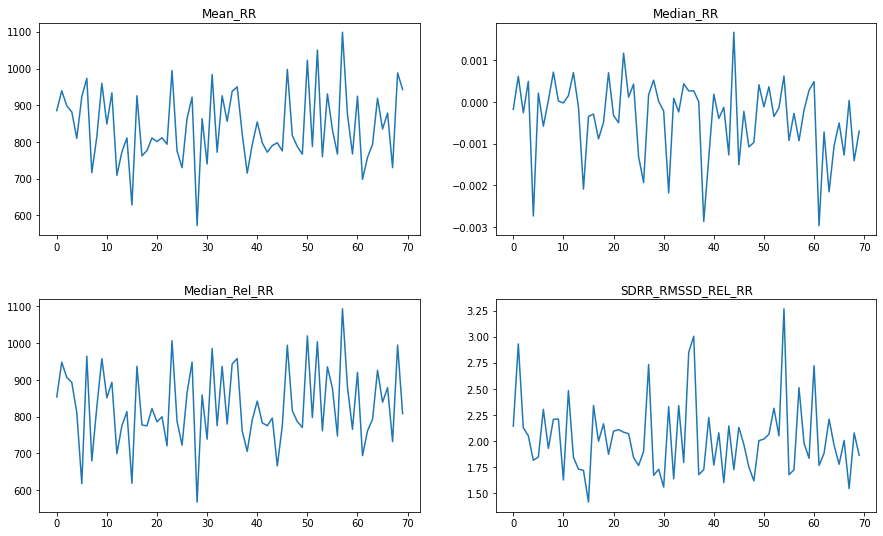
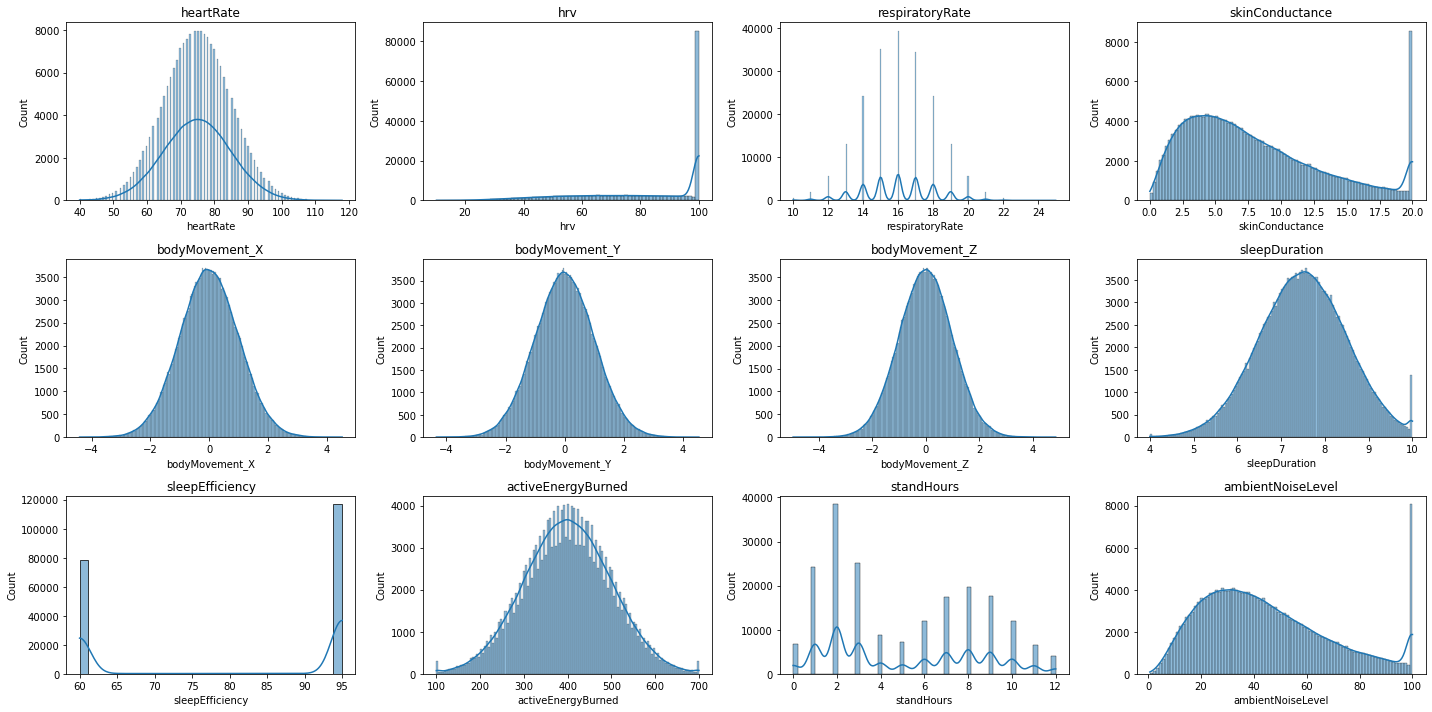
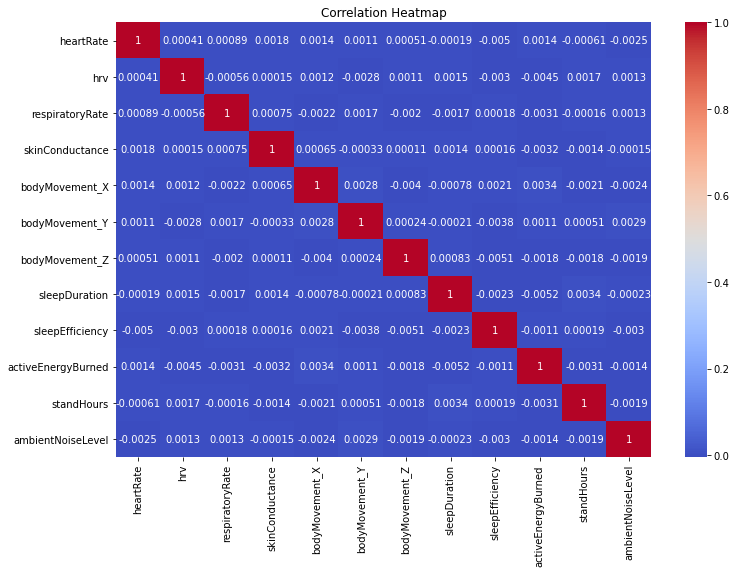
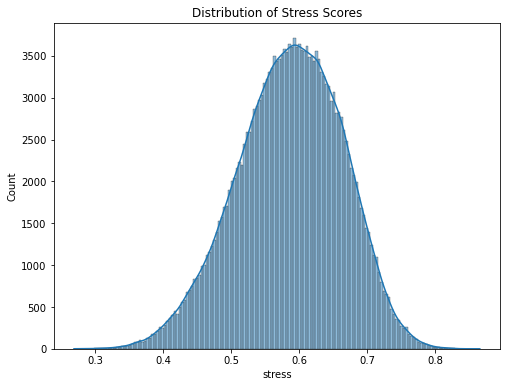


Figure 3: Plotting some of the extracted features

**Wearable Device Simulated Data**

The distribution of the synthetic data generated can be visualized below  
Figure 4: Distribution of the generated dataset for the features

  
Figure 5: Heatmap of the generated data

The distribution of the calculated stress score is also shown below.  
Figure 6: Distribution of Stress scores

The chart provides a visual representation of stress scores observed in the dataset, revealing a clear pattern in the distribution. The majority of stress scores cluster within the range of 0.5 to 0.7, indicating a prevalent level of moderate stress among the subjects. The stress scores' concentration in this range suggests that the participants experienced a consistent and relatively moderate level of stress. It's noteworthy that the chart exhibits a relatively smaller number of data points with stress scores outside the 0.5 to 0.7 range. This suggests that extreme stress levels, both high and low, were less frequently observed.