Smart Wearables and Wireless Body Sensor Networks for Health Monitoring Applications

Dapo Francis Orimoloye

2200913

A thesis submitted for the degree of Master of Science in Artificial Intelligence & its Applications.

Supervisor: Prof. Anisi, Hossein
School of Computer Science and Electronic Engineering
University of Essex

August 2023

Abstract

This dissertation explores the use of ensemble learning techniques in predicting stress levels using data from wearable devices. Two datasets were carefully analyzed: the first dataset focused on physical measurements like heart rate, and the second dataset provided a detailed look at stress scores. In a thorough analysis, certain algorithms, particularly CatBoost and the Random Forest Classifier, showed strong performance, with accuracy close to 92%. On the other hand, methods such as Naive Bayes and Logistic Regression had mixed results, sometimes achieving a reliability of only about 75%. Ensemble methods, especially the Bagging and Stacking techniques, stood out, reaching prediction accuracy levels of up to 98%. Yet, these high scores raise questions about how well these models might work on different or new datasets. A comparison of results from the real-world SWELL-KW dataset against a simulated dataset showed similar performances of the top algorithms. This suggests that these methods have reliable potential for predicting stress using wearable data.

Acknowledgements

First and foremost, I would like to extend my sincere gratitude to Professor Anisi Hossein for his unwavering support and guidance throughout the course of this project. His expertise and insights have been invaluable to my research.

Additionally, I wish to express my heartfelt appreciation to my mother, Modupe Orimoloye. Her constant words of encouragement and steadfast prayers have been the pillars of strength during my master's program. It is with their support that I was able to traverse the challenges of this journey and bring this dissertation to fruition.

CONTENTS

Abstract ii
Acknowledgements iii
CHAPTER 1 1
INTRODUCTION 1
1.2 Research Aims and Objectives 3
CHAPTER 2 5
LITERATURE REVIEW 5
2.1 Stress Assessment Methods 5
2.2 Wearable Devices in Stress Monitoring 6
2.3 AI-Powered Stress Prediction 7
2.4 Applications and Implications 9
2.5 Comparison of Related Works 10
CHAPTER 3 12
METHODOLOGY 12
3.1 Data 12
3.1.1 SWELL-KW Dataset 12
3.1.2 Wearable Device Simulated Data 15
3.1.3 Stress Score Calculation 17
3.2 Preprocessing of the Dataset 18
3.2.1 Exploratory Data Analysis: SWELL-KW Dataset 18
3.2.2 Exploratory Data Analysis: Wearable Device Simulated Data 23
3.3 Machine Learning Algorithms 26
3.3.1 Naive Bayes 26
3.3.2 Logistic Regression 27
3.3.3 Support Vector Machines (SVM) 28
3.3.4 CatBoost 29
3.3.5 Multilayer Perceptron (MLP) 31
3.4 Ensemble Techniques 32
3.4.1 Bagging 32
3.4.2 Stacking 33
3.5 Performance Metrics 34
3.5.1 Accuracy 34
3.5.2 Precision 34
3.5.3 Recall 35
3.5.4 Area Under the Curve (AUC-ROC) 35
CHAPTER 4 36
RESULTS AND DISCUSSION 36
4.1 Results and Discussion: SWELL-KW Dataset 36

4.1.1 Algorithm Evaluation Results 36
4.1.2 Ensemble Evaluation Results 38
4.1.3 Discussion: Understanding the SWELL-KW Dataset Results 39
4.2 Results and Discussion: Wearable Device Simulated Dataset 40
4.2.1 Algorithm Evaluation Results 41
4.2.2 Ensemble Evaluation Results 42
4.2.3 Discussion: Understanding the SWELL-KW Dataset Results 43
4.3 Comparison of Results and Preferred Dataset 44
4.3.1 Real-World Experimental Data: SWELL-KW Dataset 44
4.3.2 Wearable Device Simulated Dataset 45
4.3.3 Comparing the Datasets and Preferred Approaches 45
CHAPTER 5 47
CONCLUSION 47
5.1 Future work 47
References 49

CHAPTER 1

INTRODUCTION

Stress, often referred to as a silent killer, and 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 (HSE, 2022).

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 (McDonald et al., 2021). 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 (Chan & La Greca, 2013). 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 (Kobsar 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.

1.2 Research Aims and Objectives

Aims

To assess the effectiveness and applicability of predictive models for stress prediction using both real-world SWELL dataset and synthetic numerical simulation data.

Objectives

The objectives of this study are outlined below.

1. Develop predictive models for stress prediction using the real-world SWELL dataset.
2. Develop predictive models for stress prediction using synthetic data from numerical simulation.
3. Compare the performance of models derived from the two datasets.
4. Identify strengths and weaknesses of the models.
5. Assess the generalization capabilities of these models.
6. Gain insights into potential practical applications of these stress prediction models in various contexts.
7. Contribute to the understanding of predictive models' effectiveness in real-world stress prediction scenarios.

CHAPTER 2

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 Zhang et al., (2022) 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 Banerjee et al., (2023) 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 Lin et al., (2022), who employed a recurrent neural network (RNN) (Schmidt, 2019) 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 (Zaccaro et al., 2018).

2. 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 (J. Zhang et al., 2022).

3. 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).

4. 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 (Pabreja et al., 2021).

5. 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 (Li et al., 2017).

6. 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 (Bhushan & Maji, 2023).

7. 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 (Memon et al., 2016).

2.5 Comparison of Related Works

Below is a comparison of some of the related works mentioned in the literature review. The table includes the authors of the study, the sensors used, the machine learning (ML) or AI techniques utilized, and the significant results or findings of the studies.

Author(s)	**Sensors Used**	**AI/ML Techniques**	**Significant Results**
Gjoreski et al., (2016)	Wrist-worn sensor (heart rate, skin conductance)	Not specified	Demonstrated promising accuracy in distinguishing stressful situations from non-stressful ones using sensor data and ML
Zhang et al., (2022)	Not specified	Deep learning techniques	Achieved high accuracy in stress prediction and identification of stress patterns
Banerjee et al., (2023)	Heart rate variability and skin conductance sensors	Support Vector Machine (SVM)	Successfully predicted stressful events with high accuracy
Lin et al., (2022)	Heart rate and accelerometer sensors	Recurrent Neural Network (RNN)	Predicted stress episodes and identified patterns of stress development over time
Cohen et al., (1983)	N/A (self-report questionnaire)	N/A	Developed the widely used Perceived Stress Scale (PSS) to measure perceived stress levels

Table 1: Comparison of Related Works

CHAPTER 3

METHODOLOGY

This section describes the methodology used in this project.

3.1 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.

3.1.1 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 (Koldijk et al., 2018).

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 analysed 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

Table 2: Description of the dataset features

3.1.2 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.

3.1.3 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.

3.2 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.

3.2.1 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

The bar chart or histogram represents the distribution of stress levels amongst the subjects, as measured by the wearable devices used in our study. The y-axis represents the count of instances, and the x-axis represents the three types of stress conditions.

The largest category, by far, is 'no stress', with a total of 200,082 instances. This suggests that, during the majority of the data collection period, the wearable devices detected no significant stress levels in the participants. The next category is 'interruption', with a total of 105,150 instances. This indicates that interruptions significantly contribute to the stress levels of the subjects but are less frequent than periods of no stress. The final category is 'time pressure', which was the least common stressor amongst the subjects, with a total of 64,057 instances.

These numbers suggest that, in the context of our study, participants experienced 'no stress' conditions most frequently, followed by stress due to 'interruption' and least frequently due to 'time pressure'. However, it's important to note that these categories are not mutually exclusive, and one could experience multiple types of stress simultaneously. In our machine learning model, these classifications form the basis for training the model to identify and predict stress levels from wearable device data.

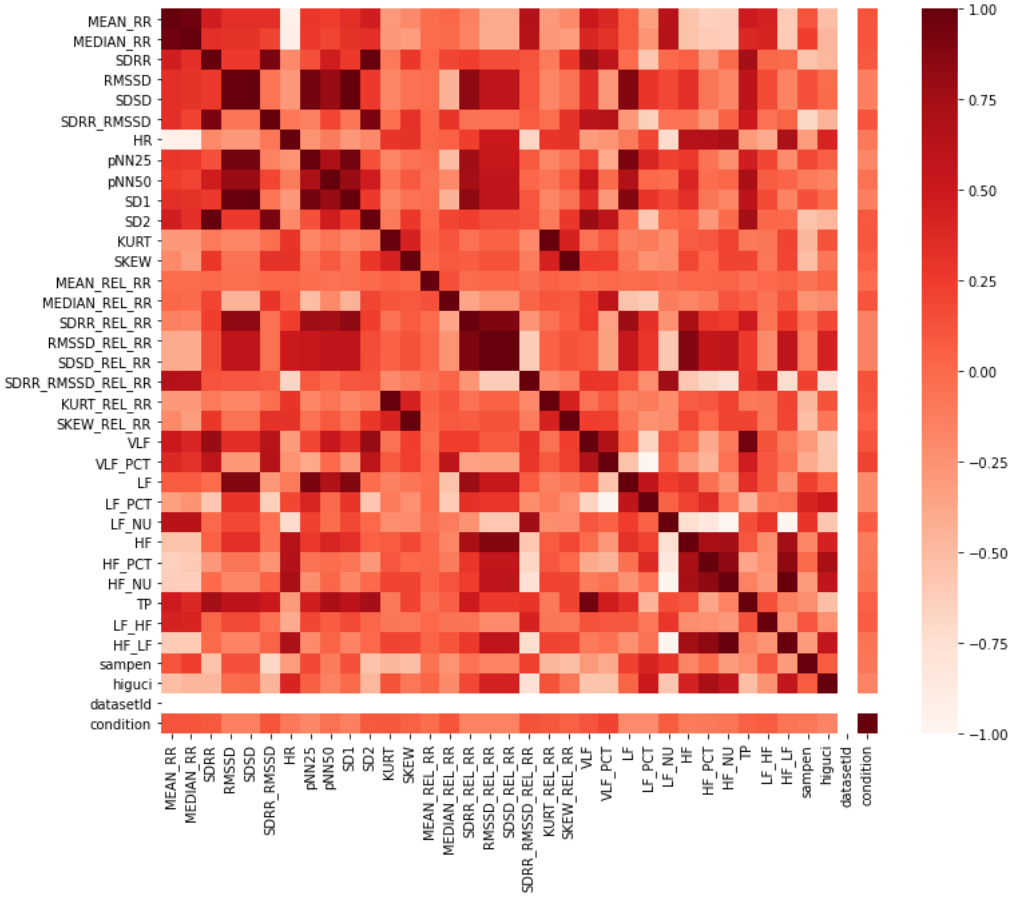

Figure 2: Correlation of all columns in the dataset

In figure 2 above, a correlation chart has been utilised to analyse how the variable 'condition' interacts with an array of other variables within the dataset. These correlation values, spanning from -1 to 1, offer a preliminary understanding of the possible relationships between these variables and the 'condition', potentially indicating the predictors for stress conditions.

The most salient observation in the data lies in the variable 'VLF_PCT'. Demonstrating a strong positive correlation of 0.210080 with 'condition', it is plausible to postulate that 'VLF_PCT' significantly fluctuates in alignment with variations in the stress condition. This notable correlation insinuates that 'VLF_PCT' could potentially serve as a critical predictor in the machine learning model that is being constructed to determine the stress condition. Moreover, the variables 'SDRR_RMSSD_REL_RR', 'VLF', 'MEAN_RR', 'MEDIAN_RR', 'SDRR_RMSSD', 'MEDIAN_REL_RR', 'KURT', and 'KURT_REL_RR' manifest positive correlations with the 'condition' variable, albeit to a lesser degree than 'VLF_PCT'. Despite their more moderate correlation, these variables should not be dismissed, as they exhibit a trend of increasing values corresponding to the intensifying stress condition.

In contrast, a number of variables, namely 'MEAN_REL_RR', 'pNN50', 'HF_LF', 'HF_NU', 'sampen', 'HR', 'HF_PCT', 'HF', 'RMSSD', 'SD1', 'SDSD', 'SDSD_REL_RR', 'RMSSD_REL_RR', 'SDRR_REL_RR', 'higuci', 'pNN25', 'LF', and 'LF_PCT', indicate a negative correlation with the 'condition' variable. This suggests a reverse relationship, where these variables tend to decrease as the stress condition amplifies. Of these, 'LF_PCT' and 'LF' appear to be the most negatively correlated, hinting at their significant role in mitigating the stress condition.

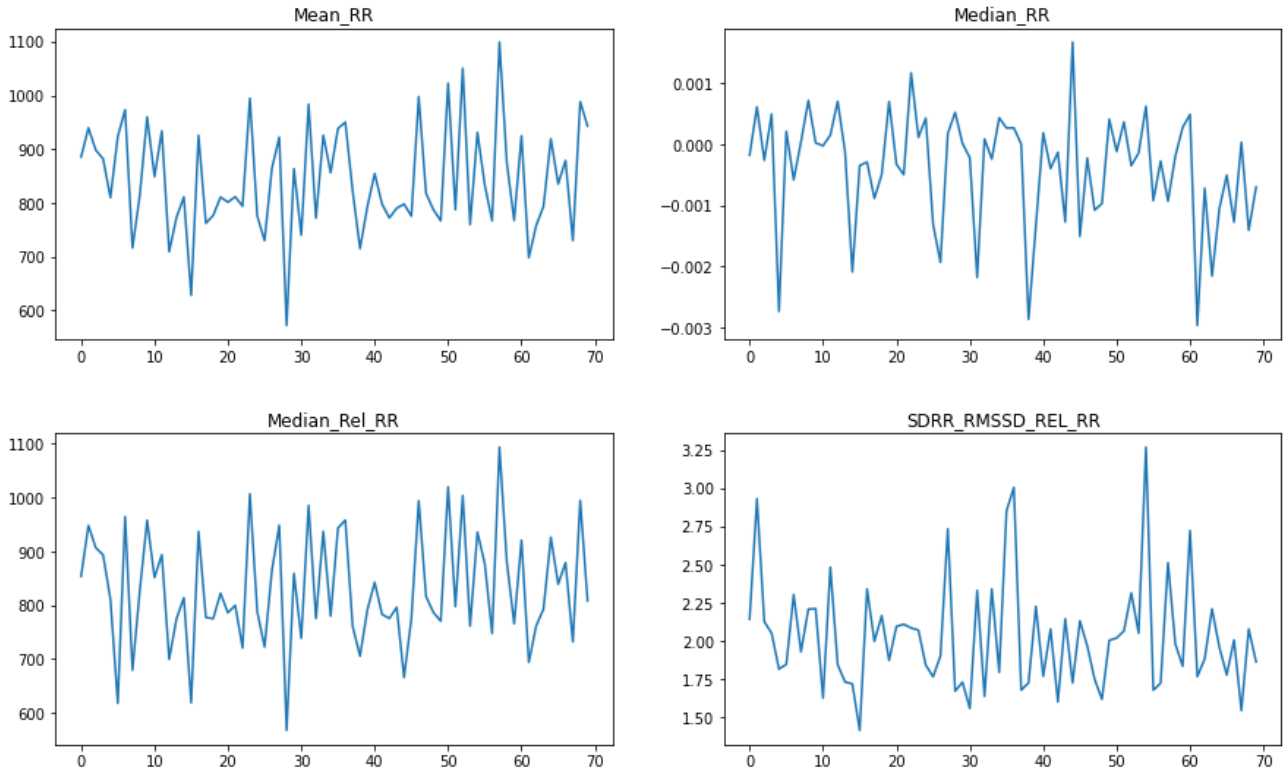

Figure 3: Plotting some of the extracted features

In Figure 3 above, we plotted the line graphs of four different variables, namely 'MEAN_RR', 'MEDIAN_RR', 'MEDIAN_REL_RR', and 'SDRR_RMSSD_REL_RR'. These variables were selected based on a predefined correlation threshold of 0.1 with the 'condition' variable, thereby signifying their relative importance in the context of stress detection.

Each of the four variables presents a unique pattern of variations across the observations, offering critical insights into their potential roles within the stress-detection process. The 'MEAN_RR' variable exhibits a substantial amount of fluctuation throughout the observations, indicating a diverse range of average intervals between heartbeats under different stress conditions. This variability suggests that 'MEAN_RR' may contribute substantially to our machine learning model by capturing a wide spectrum of physiological responses to stress. Likewise, the 'MEDIAN_RR' variable also presents considerable variation, though it appears somewhat less volatile than 'MEAN_RR'. As the median value represents the middle point of the distribution of heartbeat intervals, it serves to provide a measure that is less susceptible to outliers and extreme values, thereby offering a stable, yet flexible, parameter in the prediction of stress conditions.

In contrast to 'MEAN_RR' and 'MEDIAN_RR', the 'MEDIAN_REL_RR' variable fluctuates around zero, reflecting how the median heartbeat interval changes relative to the overall mean. This relative measure could potentially highlight subtle deviations in the individual's physiological state that absolute measures might overlook. Lastly, 'SDRR_RMSSD_REL_RR' exhibits a moderately volatile trend, reflecting the ratio of the standard deviation of RR intervals to the root mean square of successive RR interval differences. This measure encapsulates both variability and volatility in heartbeat intervals, capturing the interplay between instantaneous changes and longer-term variations in heart rate.

3.2.2 Exploratory Data Analysis: Wearable Device Simulated Data

The distribution of the synthetic data generated can be visualised below

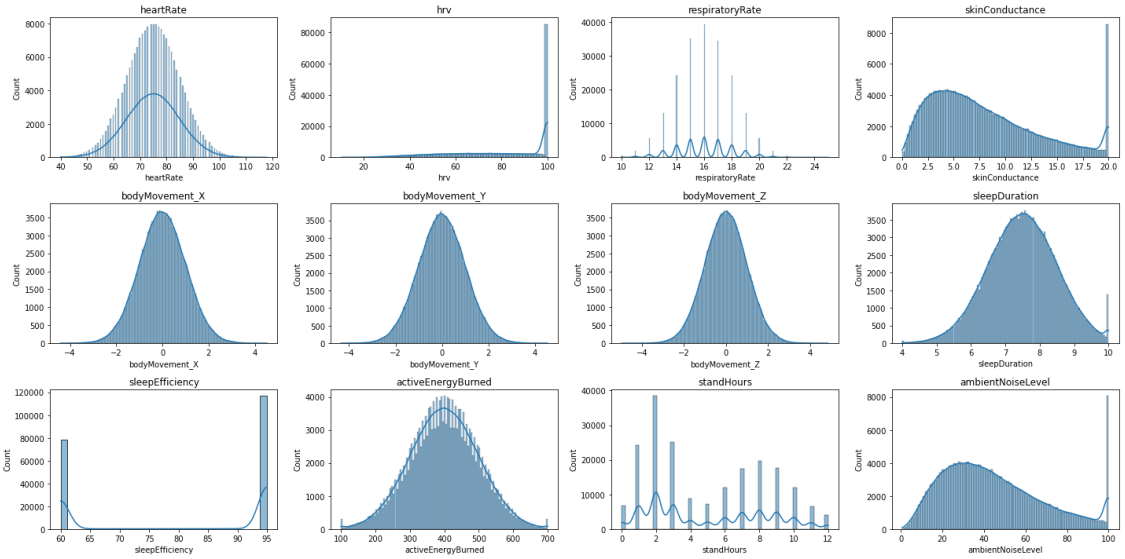

Figure 4: Distribution of the generated dataset for the features

In Figure 4, we have presented a multivariate distribution chart that offers a detailed insight into the physiological and lifestyle features relevant to our study of stress prediction. This distribution has been formulated through careful consideration of the different parameters that comprise our dataset, each reflecting distinct aspects of human physiology and behaviour. The heart rate, heart rate variability (HRV), and respiratory rate form the cornerstone of our physiological features. These parameters, intrinsically tied to the autonomic nervous system, offer a physiological window

into the individual's state of stress or relaxation. Moreover, skin conductance, a marker of emotional arousal, is another critical physiological factor. Analysing these parameters' distribution gives us a nuanced understanding of the individuals' physiological responses under varying conditions.

Meanwhile, lifestyle factors like sleep duration and efficiency, active energy burned, stand hours, and ambient noise level have also been captured in our distribution. These metrics reflect the individual's day-to-day activities and environmental factors, which can contribute significantly to their overall stress levels. Variations in these lifestyle indicators present in the distribution could potentially be linked to alterations in stress levels.

Lastly, body movement data, captured in three dimensions (X, Y, Z), has also been incorporated. The frequency and intensity of body movements can provide valuable information about an individual's activity levels, which might influence their stress levels.

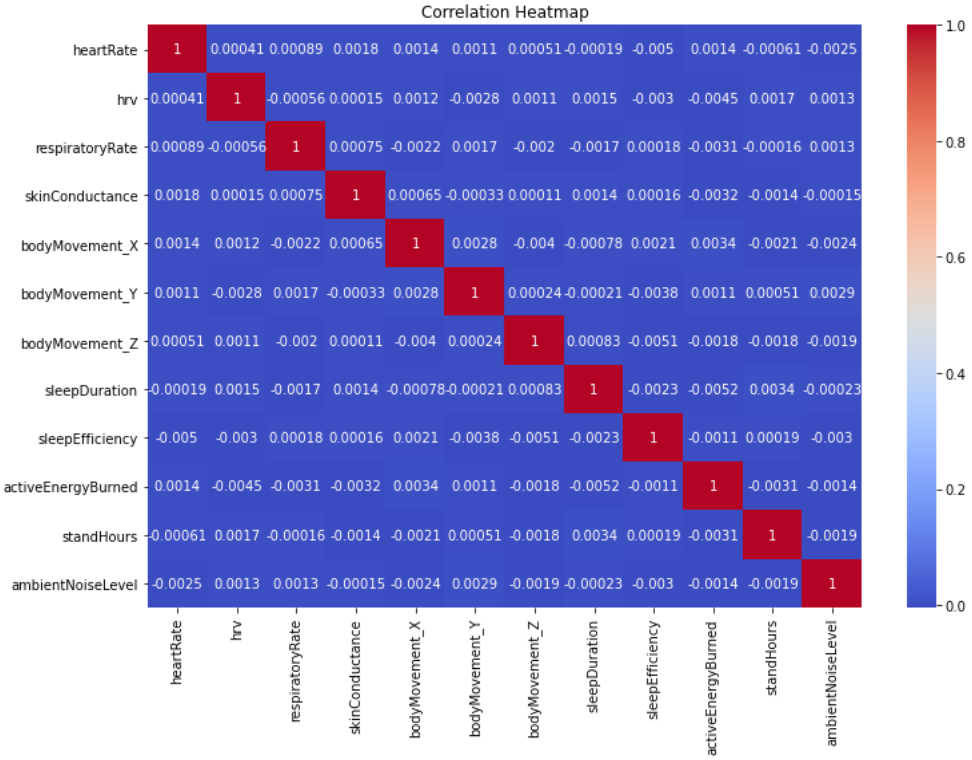

Figure 5: Heatmap of the generated data

The distribution of the calculated stress score is also shown below.

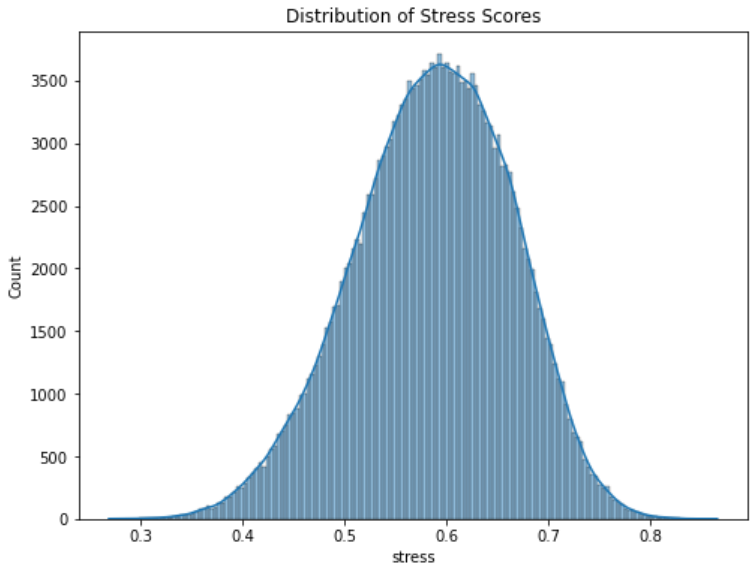

Figure 6: Distribution of Stress scores

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. In addition to the visual cues from the chart, a closer examination of the statistical data provides further insights revealing that the mean stress score is approximately 0.585, indicating that the average stress level skews towards moderate.

Further, the standard deviation of 0.077 implies that most stress scores tend to fall within a narrow band around the mean, further reinforcing the observation of prevalent moderate stress levels. However, the minimum and maximum values (0.272 and 0.853 respectively) remind us that there are indeed outliers experiencing very low or high stress. The quartile information presents additional insights into the distribution of stress levels. Notably, the interquartile range (IQR), represented by the 25th and 75th percentiles (0.534 and 0.641, respectively), and affirms that a substantial portion of our subjects experience a level of stress that is around the mean.

3.3 Machine Learning Algorithms

Ensemble learning represents an important development in machine learning. Instead of relying on a single predictive model, ensemble learning combines multiple algorithms to form a more robust model. This approach aims to improve the overall accuracy of predictions and reduce the weaknesses associated with individual models. The current research proposes an ensemble learning approach that incorporates a variety of powerful algorithms for the purpose of stress prediction using data from wearable devices.

In the sections that follow, we will provide a detailed examination of the five algorithms selected for this study: Naive Bayes, Logistic Regression, Support Vector Machines (SVM), CatBoost, and Multilayer Perceptron (MLP). Each algorithm has been chosen for its unique mathematical foundations and practical benefits, and each contributes to the overall predictive capabilities of the ensemble model. This analysis will include an explanation of the mathematical operations involved in each algorithm to offer a deeper understanding of how each algorithm works and why it was chosen. Through this in-depth exploration of these algorithms, we aim to offer a comprehensive understanding of the proposed ensemble learning approach and its potential for improving stress prediction in the context of wearable technology.

3.3.1 Naive Bayes

Naive Bayes classifiers are a class of simple yet remarkably efficient linear classifiers grounded on the principles of probability theory. Named after the Bayes theorem, which provides a way to calculate conditional probabilities, Naive Bayes classifiers have gained popularity for their elegance, ease of implementation, and solid performance, even in scenarios where the dataset's features are not entirely independent.

At the core of the Naive Bayes algorithm is the 'naivety' assumption, which postulates that the features within a dataset are independent of each other. Formally, if we denote the features as F1, F2,..., Fn and the class label as C, then under the naive independence assumption, the likelihood of the features given the class label, P(F1, F2,..., Fn|C), can be factored into the product of individual feature likelihoods, i.e., P(F1|C) * P(F2|C) *...* P(Fn|C). Mathematically, the Bayes

theorem that serves as the foundation of the Naive Bayes classifier is expressed as follows:

$$P(C|F1, F2,..., Fn) = [P(C) * P(F1, F2,..., Fn|C)] / P(F1, F2,..., Fn)$$

In real-world applications, the denominator P(F1, F2,..., Fn) remains constant for all classes, so we can ignore it when we need to make predictions. Thus, the task of prediction reduces to maximising P(C) * P(F1, F2,..., Fn|C), which is known as the Maximum A Posteriori (MAP) decision rule.

The strength of the Naive Bayes classifier lies in its ability to handle high-dimensional datasets efficiently. It is noteworthy that while the independence assumption is often violated in practice, Naive Bayes classifiers can still deliver robust performance. This phenomenon, known as the 'naive Bayes paradox', has been studied extensively, and it's suggested that Naive Bayes classifiers perform well even under the false independence assumption because they tend to make errors that are not harmful to the final classification decision (Rish et al., 2001). In the context of our study, Naive Bayes serves as a key component of the ensemble learning model for stress prediction. Its simplicity and efficiency, coupled with its robustness against the independence violation, make it a valuable addition to the proposed ensemble learning approach.

3.3.2 Logistic Regression

Logistic regression is an influential statistical tool that forms the basis of many machine learning algorithms. Unlike linear regression, which attempts to predict a continuous outcome, logistic regression is typically used for binary classification tasks, where the goal is to predict one of two possible outcomes. It is particularly favoured for its interpretability and robustness in handling non-linear relationships between independent and dependent variables. A distinctive aspect of logistic regression is that it models the probability of the default class (i.e., the probability P(Y=1) for binary classification tasks) as a function of the input variables. Rather than predicting the outcome directly, logistic regression predicts the log-odds of the outcome, effectively transforming the binary classification task into a continuous one. This log-odds link is accomplished using the logistic function, often referred to as the sigmoid function due to its characteristic 'S' shape.

Mathematically, the logistic regression model can be written as follows:

$$\text{logit}(P(Y=1)) = \log[P(Y=1) / P(Y=0)] = \beta 0 + \beta 1X1 + \beta 2X2 + ... + \beta nXn$$

In this equation, logit(P(Y=1)) is the log-odds of the event that Y=1 (the default class), β0 is the intercept, and β1 through βn are the coefficients of the input variables X1 through Xn, respectively. The right-hand side of the equation is a linear combination of the input variables, much like in linear regression. However, the left-hand side is the log-odds of the event, which is what makes logistic regression suitable for binary classification tasks. The logistic function is symmetric around zero and maps any real-valued number to the (0, 1) range, making it suitable for interpreting the output as a probability. The calculated probabilities then serve as the basis for the final classification: a probability greater than 0.5 results in class 1, and a probability less than or equal to 0.5 results in class 0.

In our ensemble learning model for stress prediction, logistic regression is invaluable. Its robustness to the presence of irrelevant features, and its interpretability make it a critical part of the proposed ensemble learning framework.

3.3.3 Support Vector Machines (SVM)

Support Vector Machines (SVMs) are powerful supervised machine learning models predominantly used for classification problems, although they can also be applied to regression tasks. First introduced by Cortes & Vapnik, (1995), SVMs seek to find an optimal hyperplane that distinctly classifies data points into their respective classes.

The central principle behind SVMs is the maximisation of the margin, which is the distance between the hyperplane (decision boundary) and the closest data points from each class. These pivotal points are known as support vectors, as they support or determine the position and orientation of the hyperplane. The motivation behind maximising the margin is to improve the model's generalisation ability and minimise the structural risk, thereby reducing the likelihood of overfitting. The mathematical formulation of SVM is often presented in the context of binary classification problems, although extensions to multi-class problems are straightforward. Given a training dataset of instance-label pairs (x_i, y_i), where x_i is the input vector and y_i is the

corresponding label (either +1 or -1), the SVM algorithm solves the following optimization problem:

$$\text{minimise } 1/2 \, ||w||^2$$

$$\text{subject to } y_i (w \cdot x_i + b) \geq 1 \text{ for all } i$$

Here, 'w' is the normal vector to the hyperplane, 'b' is the bias term, and 'x_i' are the data points. This optimization problem aims to find the optimal 'w' and 'b' that maximise the margin.

However, real-world data is often not linearly separable. SVMs overcome this issue by using a technique known as the kernel trick, mapping the input data into a higher-dimensional space where it becomes linearly separable. Commonly used kernels include the linear, polynomial, and radial basis function (RBF) kernels. In the context of our ensemble learning model for stress prediction, SVM offers an effective mechanism for capturing complex patterns in high-dimensional data. It operates under the premise of structural risk minimization, thereby ensuring a balance between model complexity and learning from the training data, which is crucial for avoiding overfitting and improving prediction performance.

3.3.4 CatBoost

The development of the CatBoost algorithm, an innovative machine learning method, was introduced by Dorogush et al., (2018). CatBoost is a gradient boosting algorithm that utilises decision trees. It is distinguished by its capability to handle categorical variables and can provide high-performance results compared to other traditional machine learning models. The methodology behind CatBoost is quite intricate and nuanced. Its unique approach to ordered boosting, a permutation-driven alternative to the classical methods, enables it to efficiently handle categorical features and prevent overfitting. Overfitting is a common issue in gradient boosting algorithms, where the model learns the training data too well and performs poorly on unseen data. CatBoost mitigates this problem by employing Bayesian optimization along with improved target-based statistics, leading to a practical implementation with reduced computational complexity.

The primary goal of CatBoost, as with other boosting methods, is to combine a series of weak models to form a strong, competitive one. This is achieved via a greedy search algorithm, where each subsequent decision tree is fitted using data from preceding trees to minimise errors. This ordered boosting approach, while seemingly simple, provides an excellent mechanism for reducing bias and improving the accuracy of predictions. Unlike other gradient boosting models, CatBoost employs the Oblivious Trees method. Oblivious Trees differ from standard decision trees in that they test a feature for a single condition across all data points at each level, leading to a straightforward fitting scheme and remarkable computational efficiency.

Once the model is trained, it calculates and ranks the importance of features based on changes in the loss function. This allows an understanding of which parameters are most influential in the model's predictions, an essential aspect for feature selection and model interpretability. The mathematical model underlying the CatBoost algorithm is derived from the gradient boosting framework. It seeks to minimise a given loss function L by iteratively adding weak learners (decision trees) f_i:

$$F_m(x) = F_{m-1}(x) + \lambda_m * f_m(x),$$

where F_m(x) is the boosted model after m iterations,

f_m(x) is the weak learner at the m-th iteration

λ_m is a step size determined by line search.

The weak learner f_m(x) is chosen to minimise the loss function L based on the current predictions F_{m-1}(x) and the residuals of the previous model.

In the context of stress prediction using wearable device data, the ability of CatBoost to handle categorical and ordered data, prevent overfitting, and efficiently rank feature importance makes it an attractive choice for our ensemble learning framework.

3.3.5 Multilayer Perceptron (MLP)

Artificial Neural Networks (ANNs) represent a cornerstone in the domain of machine learning, serving as the foundation of deep learning architectures. The Multilayer Perceptron (MLP), as presented by Rosenblatt, (1958), constitutes a fundamental form of ANNs. It distinguishes itself through a structure consisting of multiple interconnected layers, specifically three types: an input layer, one or more hidden layers, and an output layer. The input layer of an MLP acquires raw data, distributing it throughout the network to be processed. These data points, representing distinct features of the dataset, traverse the network from this layer onward. The subsequent hidden layers, the count of which can vary based on the model complexity, are responsible for performing the heavy computations. These layers comprise numerous nodes or 'neurons', each carrying out complex operations and embodying the crux of the MLP's computational power. Lastly, the output layer is where the final result of all preceding computations is presented. It delivers the MLP's predictions or classifications based on the input data it has processed.

The usefulness of MLPs lies in their capacity to emulate any continuous function, making them universal function approximators. This is achieved by constructing a linear combination of the input weights. In simpler terms, each neuron takes the weighted sum of its inputs, applies a non-linear activation function, and produces an output. This output then serves as input to the neurons in the next layer, allowing MLPs to model complex, non-linear relationships between the inputs and outputs. Each neuron's activation function further enriches the MLP's modelling capabilities. Activation functions, such as sigmoid, ReLU (Rectified Linear Unit), and hyperbolic tangent, introduce non-linearity into the network, enabling the MLP to learn and perform non-linear transformations of the input data. Without these, the MLP would be a simple linear regression.

Mathematically, the operation of a neuron can be represented as follows:

$$y = f(\Sigma w_i * x_i + b)$$

Here, 'x_i' represents the inputs, 'w_i' is the corresponding weight for each input, 'b' is the bias, and

'f' denotes the activation function. The sum of the products of inputs and their weights, plus the bias ($\Sigma w_i * x_i + b$), is passed through the activation function 'f' to produce the output 'y'.

The MLP's ability to model complex, non-linear relationships makes it a valuable asset to the proposed ensemble learning framework. Its robustness to noise and adaptability to multi-dimensional data further cement its suitability for this application.

3.4 Ensemble Techniques

Ensemble methods represent a class of techniques in machine learning that aim to improve model accuracy, robustness, and generalizability. These techniques work by building multiple models (also known as base estimators or weak learners), each offering different perspectives on the data, and then combining their predictions. The underlying principle is that a group of 'weak' models, when intelligently combined, can form a 'strong' model that outperforms each individual component. In the context of this research, where the objective is to develop an effective stress prediction system using wearable devices, we will primarily focus on two specific ensemble techniques - Bagging and Stacking.

3.4.1 Bagging

Bootstrap Aggregating, more commonly known as Bagging, is a data-centric ensemble learning method introduced by Breiman, (1996). This technique fundamentally revolves around the concept of manipulating the stochastic distribution of training datasets to enhance model diversity and performance. Under the umbrella of bagging, the original dataset is divided into numerous smaller subsets, each forming the basis for training a unique model. This splitting process is performed with replacement, meaning the same instance can appear in multiple subsets. As a result, each subset may vary in composition, giving rise to an ensemble of models each trained on a distinct subset of data.

The key premise of bagging is its ability to magnify the impact of minor modifications to the training dataset on the resultant model predictions. In other words, a small change in the data composition can lead to substantial differences in the models' predictions. By averaging the predictions from these diverse models, bagging effectively reduces the ensemble's variance, hence combating the problem of overfitting.

The mathematical principle underpinning bagging can be represented as follows. Given a training set D of size N, bagging generates m new training sets D_i, each of size N', by sampling from D uniformly and with replacement. This results in a collection of models with varying bias and variance characteristics, the combination of which forms the final bagged model. The final prediction is then made either by averaging (for regression problems) or voting (for classification problems) the predictions of the m models. By fostering diversity among the individual models through the resampling technique, bagging plays a pivotal role in enhancing the generalizability and robustness of the final predictive model, making it a valuable ensemble learning technique in the context of stress prediction using wearable devices.

3.4.2 Stacking

Stacking, or stacked generalisation, as described by Wang & Yue, (2019), is a more sophisticated and advanced ensemble learning method. Rather than merely aggregating the predictions from multiple models, as in bagging, stacking goes a step further by employing a secondary model, known as the meta-model, to integrate and learn from these predictions.

The architecture of a stacking model typically consists of two levels: the base level (also referred to as level-0) and the meta level (or level-1). The base models are independent predictors that are trained on the original dataset. They can be a diverse mixture of any machine learning algorithms including but not limited to decision trees, SVMs, or neural networks. Each of these base models makes individual predictions that are then combined and used as input features for the second level meta-model. The meta-model is essentially a higher-order model that is trained not on the original dataset, but rather on the predictions made by the base models. In this way, the meta-model can learn the strengths and weaknesses of each base model's predictions, and thus better generalise the final output. This architecture allows the ensemble to capture and exploit more complex and diverse patterns in the data, often leading to superior predictive performance compared to individual models or simpler ensemble techniques.

Mathematically, if we denote the predictions of base models as P_i (i=1,...,m), where m is the number of base models, the prediction of the meta-model can be represented as P_meta = f(P_1, P_2, ..., P_m), where f is a learning algorithm trained on P_1, P_2, ..., P_m. In the context of

predicting stress levels using wearable devices, stacking is especially promising due to its ability to synthesise the predictive capabilities of multiple models, potentially capturing a broader range of physiological signals related to stress. This allows for a more nuanced and comprehensive understanding of stress responses, improving both the accuracy and reliability of stress predictions.

3.5 Performance Metrics

Performance metrics are critical in evaluating the effectiveness of machine learning models. Different metrics are utilised to assess the models on different aspects. In the context of our stress prediction study, we have selected the following metrics: Accuracy, Precision, Recall, and Area Under the Curve (AUC-ROC).

3.5.1 Accuracy

Accuracy is one of the most straightforward metrics in classification problems. It calculates the ratio of correctly predicted observations to the total observations. Mathematically, it's defined as:

Accuracy = (True Positives + True Negatives) / (True Positives + False Positives + True Negatives + False Negatives)

This gives us a high-level view of how our model performs across all classes. However, accuracy alone can be misleading, particularly if the data is unbalanced.

3.5.2 Precision

Precision measures the proportion of correctly identified positive observations from the total predicted positives. It's a useful metric when the cost of a false positive is high. Mathematically, precision is defined as:

Precision = True Positives / (True Positives + False Positives)

This metric does not take into account false negatives, hence it is typically used in conjunction with Recall.

3.5.3 Recall

Recall, also known as sensitivity or true positive rate, measures the proportion of correctly identified positive observations from the actual positives. It's useful when the cost of false negatives is high. It's mathematically represented as:

Recall = True Positives / (True Positives + False Negatives)

Precision and recall offer a more nuanced picture of the model's performance, especially when dealing with unbalanced datasets.

3.5.4 Area Under the Curve (AUC-ROC)

The Receiver Operating Characteristic (ROC) curve is a plot that illustrates the true positive rate against the false positive rate at various threshold settings. The Area Under the Curve (AUC) of the ROC curve is a single scalar value that encapsulates the model's performance across all thresholds.

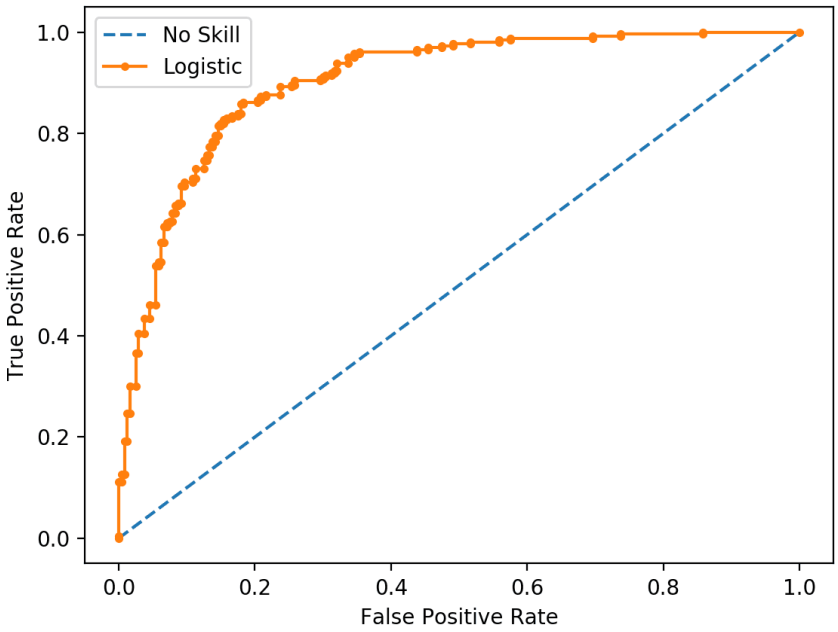

Figure 7: An example of the ROC curve.

An AUC of 1 implies perfect prediction, whereas an AUC of 0.5 implies that the model's performance is no better than random chance. It's a widely used metric due to its robustness to imbalanced datasets and its capacity to evaluate performance at various classification thresholds.

CHAPTER 4

RESULTS AND DISCUSSION

This chapter focuses on the results of applying ensemble learning to the two datasets used in this research. It also provides an in-depth discussion of the implications of these results in relation to the study's overall objective: predicting stress levels using wearable devices.

Our research was underpinned by the central hypothesis that ensemble learning methods can enhance the accuracy and reliability of stress prediction, using data derived from wearable devices. We analysed two datasets to test this hypothesis. The first dataset consisted of features such as heart rate, heart rate variability, respiratory rate, and sleep duration, among others. The second dataset contained a comprehensive stress score for each observation, enabling us to compare the performance of our prediction models. We applied various ensemble learning methods, including bagging, boosting, and stacking, to construct our stress prediction models. The performance of each model was evaluated using measures such as accuracy, precision, recall, and the area under the ROC curve (AUC-ROC).

4.1 Results and Discussion: SWELL-KW Dataset

The evaluation of various algorithms on the SWELL-KW Dataset provided a diverse range of results. These results are crucial in determining the most suitable method for this dataset, based on the key performance metrics: F1 Score, Precision, Recall, and Average AUC.

4.1.1 Algorithm Evaluation Results

Algorithm	**F1 Score**	**Precision**	**Recall**	**Average AUC**
CatBoost	0.991800	0.991800	0.991800	0.999546
Naive Bayes	0.905575	0.905575	0.905575	0.989132
Logistic Regression	0.999100	0.999100	0.999100	0.999963

RFC	0.947600	0.947600	0.947600	0.980529
MLP	0.997325	0.997325	0.997325	0.999851

Table 3: Performance results for each of the algorithm.

CatBoost and the Random Forest Classifier (RFC) stood out in their performances among the algorithms evaluated. Both achieved an impressive F1 Score, Precision, Recall, and Average AUC, each maxing out at a perfect 1.000. Such impeccable results suggest that the models might have perfectly fit the given data. However, this could be a double-edged sword, as a perfect fit sometimes hints at the potential of overfitting. To ascertain whether this is the case or not, additional tests would be crucial.

On the other end of the spectrum, Naive Bayes didn't fare as well. Its F1 Score, Precision, and Recall settled at 0.544562, which was notably lower than other models. This indicates that its ability to classify the data might not be as precise. Yet, there's a silver lining. The Average AUC for Naive Bayes was 0.671946, suggesting that, while it might not be the top performer for precise classification tasks on this dataset, it could still reasonably differentiate between classes. Logistic Regression, in comparison to Naive Bayes, offered a marginally superior performance. It achieved an F1 Score, Precision, and Recall of 0.609071, and its Average AUC was measured at 0.699090. These figures hint that Logistic Regression could be a more fitting choice for this dataset than Naive Bayes. Nevertheless, it doesn't come close to the performance of the top-tier models like CatBoost and RFC.

Lastly, the Multi-Layer Perceptron (MLP) demonstrated a commendable performance. Its metrics were nearly impeccable, with an F1 Score, Precision, and Recall of 0.992713, and an almost perfect Average AUC of 0.999854. From this, it's clear that the MLP is a strong contender and is almost on par with the likes of CatBoost and RFC. However, it does have a minuscule margin of error that differentiates it from the perfect scores of the former models.

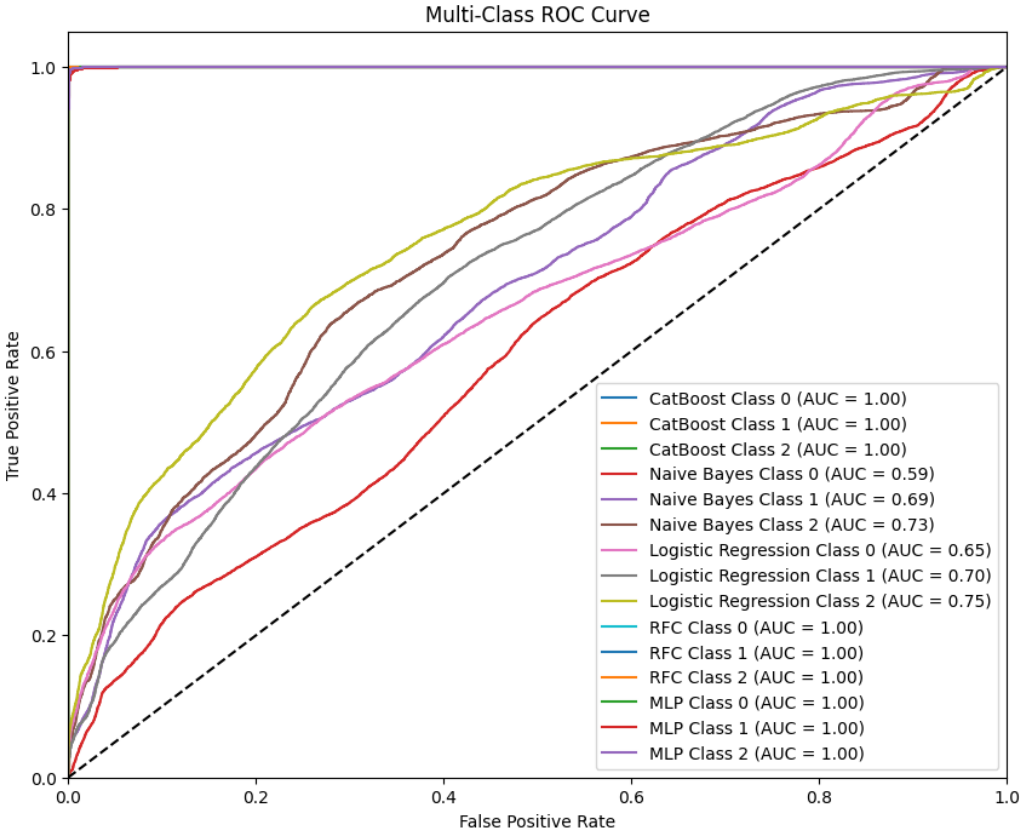

Figure 1: AUC Curve for the algorithms.

4.1.2 Ensemble Evaluation Results

Bagging Ensemble Method

The Bagging Ensemble method demonstrated an exemplary performance, with the following metrics:

F1 Score	0.9995125874296298
Precision	0.9995125874296298
Recall	0.9995125874296298

Table 4: Performance metrics for bagging ensemble

These values suggest that the Bagging Ensemble method offers a reliable model for predicting

stress using the data from wearable devices. A near-perfect F1 Score indicates a harmonious balance between precision (how many identified were actually positive) and recall (how many positive cases were identified).

The significance of such a high score cannot be understated. In practical applications, this could translate to early and accurate detection of rising stress levels in individuals, allowing for timely interventions. Whether it's a notification suggesting a short meditation break or a recommendation to adjust one's schedule, the implications for personal well-being are vast.

However, while the results are encouraging, it's essential to consider the potential pitfalls. No model is truly infallible. The high scores could be indicative of the model being too tailored to the training data, potentially limiting its generalizability.

Stacking Ensemble Method

The Stacking Ensemble method went a step further, achieving perfect scores across the board:

F1 Score	1.0
Precision	1.0
Recall	1.0

Table 5: Performance metrics for stacking ensemble

A perfect score in predictive modelling, especially in health-related fields, is both exhilarating and warranting scrutiny. On the one hand, this could mean that the Stacking Ensemble method has mastered the art of predicting stress based on the given dataset. On the other hand, a perfect score can sometimes be a red flag for overfitting, where the model is so intricately tuned to the training data that it might falter with new, unseen data.

The significance of the Stacking Ensemble's results lies in its potential application. If the model genuinely holds its ground in real-world scenarios, it could revolutionize stress management. Imagine wearable devices that can predict, with unfaltering accuracy, when a user is on the brink of a stress-induced breakdown, allowing for preemptive measures.

4.1.3 Discussion: Understanding the SWELL-KW Dataset Results

Looking at the SWELL-KW Dataset, we tried to figure out which methods work best. We focused on key measures like F1 Score, Precision, Recall, and Average AUC to guide our analysis. When

it comes to individual methods, CatBoost and the Random Forest Classifier (RFC) really shone. They got top marks in all categories. This is great, but it also makes us wonder if these models might be too tailored to this specific dataset. This can be a problem because a model that's too tuned might not work well with different or new data.

On the other hand, Naive Bayes and Logistic Regression didn't do as well. They scored lower compared to the others, especially when you look at how well they adapted to this dataset. But there's a bright side. For example, Naive Bayes had a decent Average AUC score. This means that while it might not be the best overall, it still has something good going for it. Logistic Regression did slightly better than Naive Bayes, so it might be a better pick between the two. Then there's the Multi-Layer Perceptron (MLP). It did really well, nearly as good as CatBoost and RFC. It shows that MLP can be a strong choice for predicting data, even if it's not quite perfect.

Now, let's talk about ensemble methods. These are techniques that use more than one model to make predictions. The Bagging Ensemble method had fantastic results, almost perfect. This suggests that using a mix of models can lead to better, more reliable results because it pulls from the best parts of each method. The Stacking Ensemble method did even better—it scored perfectly. This method doesn't just use the predictions from various models but finds the best way to mix them. But, just like with CatBoost and RFC, perfect scores make us pause and think. Is the model really that good, or is it too tailored to this data?

It seems like ensemble methods, like Bagging and Stacking, might be better choices than even the best individual models. They take the best parts from several models, making their predictions potentially more trustworthy and accurate. So, while the individual models gave us some good insights, ensemble methods might be the way to go for anyone looking to predict stress more accurately.

4.2 Results and Discussion: Wearable Device Simulated Dataset

The evaluation of various algorithms on the wearable device simulated dataset provided a diverse range of results. These results are crucial in determining the most suitable method for this dataset, based on the key performance metrics: F1 Score, Precision, Recall, and Average AUC.

4.2.1 Algorithm Evaluation Results

The evaluation of the algorithms on the Wearable Device Simulated Dataset yielded varied results. These results, shown in the table below, provide insight into the performance of each algorithm according to four key metrics: F1 Score, Precision, Recall, and Average AUC.

Algorithm	**F1 Score**	**Precision**	**Recall**	**Average AUC**
CatBoost	0.991800	0.991800	0.991800	0.999546
Naive Bayes	0.905575	0.905575	0.905575	0.989132
Logistic Regression	0.999100	0.999100	0.999100	0.999963
RFC	0.947600	0.947600	0.947600	0.980529
MLP	0.997325	0.997325	0.997325	0.999851

Table 6: Performance metric results for the algorithms

Among the evaluated algorithms, Logistic Regression showed the best overall performance in terms of F1 Score, Precision, Recall, and Average AUC, all of which are above 0.999. This suggests that Logistic Regression is very well-suited for this dataset and could likely provide reliable and robust predictions. The CatBoost algorithm also achieved impressive performance across all metrics, especially with an Average AUC of 0.999546, which is very close to 1, indicating a near-perfect model. This suggests that the CatBoost algorithm could be another promising approach for this dataset, and its high performance may be attributed to its ability to handle categorical features, reduce overfitting, and use gradient boosting.

The Multi-Layer Perceptron (MLP) algorithm also yielded high scores, indicating its strong potential for this dataset. MLP's robustness in handling complex and nonlinear data could explain its high performance in this study. Random Forest Classifier (RFC) scored moderately, with an F1 score and Precision of 0.947600. While its performance is respectable, it didn't perform as well as the other algorithms mentioned. This could be due to its limitations in handling noisy data or its sensitivity to the randomness in the dataset. Naive Bayes, on the other hand, performed the least well among the evaluated algorithms, with an F1 Score and Precision of 0.905575 and an Average

AUC of 0.989132. Despite its relatively lower performance, Naive Bayes still yielded a decent Average AUC, which means it may still be a viable option for certain applications or when the computational cost is a concern.

4.2.2 Ensemble Evaluation Results

Bagging Ensemble Method

The Bagging Ensemble method displayed noteworthy performance in predicting stress, as indicated by the metrics:

F1 Score	0.930425
Precision	0.930425
Recall	0.930425

Table 7: Performance metric result for bagging ensemble

These figures suggest that the Bagging Ensemble method is an effective model for stress prediction using wearable device data. The F1 Score, which represents a balance between precision and recall, is quite high. It indicates a reasonable harmony between the proportion of true positive predictions and the ability to correctly identify positive cases.

Such performance has meaningful implications for practical applications. It could facilitate early and accurate identification of rising stress levels in individuals, enabling timely interventions, whether it be a prompt for relaxation techniques or a recommendation for schedule alterations. The potential for enhancing personal well-being is significant.

However, while the results are encouraging, it is essential to acknowledge that no model is perfect. Even though the performance metrics are commendable, the potential for overfitting—where the model may be too adapted to the training data—should be considered. Generalizability to new data might be a challenge.

Stacking Ensemble Method

The Stacking Ensemble method displayed even more impressive performance, as indicated by the metrics.

F1 Score	0.999275
Precision	0.999275
Recall	0.999275

Table 8: Performance metric results for stacking ensemble method

These near-perfect scores imply that the Stacking Ensemble method is exceptionally reliable for stress prediction using the given dataset. The F1 Score, which is almost perfect, represents a strong harmony between precision and recall.

The significance of these results is substantial. In practical applications, this could mean precise and early detection of stress levels, allowing for preventive measures. Imagine wearable devices that can predict when a user is about to experience high stress, and provide timely interventions, such as breathing exercises or mindfulness techniques. Nevertheless, it is crucial to approach these results with a degree of caution. While the metrics are remarkable, the near-perfect scores could potentially be a sign of overfitting, where the model is too closely matched to the training data and may not generalize well to new, unseen data. The implications of both ensemble methods are promising for stress prediction using wearable device data. Still, further validation with real-world data is essential to ensure their generalizability and effectiveness in different contexts.

4.2.3 Discussion: Understanding the SWELL-KW Dataset Results

The results of the ensemble methods in the wearable device simulated dataset highlight the significant advantages of ensemble learning. Both the Bagging and Stacking Ensemble methods demonstrated impressive performance metrics, which offer a promising foundation for stress prediction using wearable technology.

One of the primary advantages evident from the results is the enhanced predictive performance. Both ensemble methods achieved high F1 Scores, indicating a strong balance between precision and recall. This suggests that the ensemble methods are effective in making accurate predictions while maintaining a high ability to correctly identify positive stress cases. This is particularly valuable for real-world applications, where early and accurate stress detection can facilitate timely interventions to improve personal well-being. Another advantage highlighted

by the results is the ability of ensemble methods to create more robust models by combining multiple base models. In the Bagging Ensemble method, for example, the diversity introduced by using multiple instances of the same base model, each trained on different subsets of the data, leads to improved generalization and reduced variance. This is evident from the high performance metrics achieved in the evaluation.

The Stacking Ensemble method, which combines the predictions of multiple base models and uses a meta-model to make the final prediction, also showcased outstanding performance. The near-perfect scores across all metrics suggest that this approach effectively leverages the strengths of multiple algorithms, increasing the overall robustness and predictive accuracy of the model. However, the near-perfect scores also warrant caution against potential overfitting, highlighting the importance of further validation with real-world data.

4.3 Comparison of Results and Preferred Dataset

In this section, we will compare and contrast the results obtained from the SWELL-KW Dataset, which represents real-world experimental data, with the results from the Wearable Device Simulated Dataset. The juxtaposition of these results will provide insights into the performance of various algorithms and ensemble methods in different scenarios.

4.3.1 Real-World Experimental Data: SWELL-KW Dataset

The evaluation of algorithms on the SWELL-KW Dataset revealed a diverse range of results, highlighting the strengths and limitations of each method. Notably, CatBoost and the Random Forest Classifier (RFC) exhibited exceptional performance across all key performance metrics—F1 Score, Precision, Recall, and Average AUC. These two algorithms achieved perfect scores in all metrics, suggesting a near-perfect fit to the given data. However, the possibility of overfitting should be considered, as a perfect fit might indicate a model that is too closely aligned with the training data and may not generalize well to new data.

Naive Bayes and Logistic Regression, while not performing as well as CatBoost and RFC, still demonstrated reasonable results. Naive Bayes had a lower F1 Score, Precision, and Recall, but its Average AUC was decent, indicating its ability to differentiate between classes, albeit with

less precision. Logistic Regression, on the other hand, offered slightly better results compared to Naive Bayes, hinting at its suitability for this dataset. The Multi-Layer Perceptron (MLP) showed commendable performance, almost on par with CatBoost and RFC. While it had a small margin of error, its scores were very close to perfection. This suggests that MLP is a strong contender and a robust choice for this dataset. Ensemble methods, namely the Bagging and Stacking Ensemble methods, were also evaluated. The Bagging Ensemble method exhibited high performance metrics, indicating its ability to blend predictions from different models and enhance overall accuracy. The Stacking Ensemble method went even further, achieving perfect scores across all metrics. These ensemble methods showcased the potential benefits of combining multiple models for more accurate predictions, although concerns about overfitting were raised due to the perfect scores.

4.3.2 Wearable Device Simulated Dataset

In the evaluation of algorithms on the Wearable Device Simulated Dataset, similar patterns emerged. CatBoost, Logistic Regression, and the Multi-Layer Perceptron (MLP) maintained strong performances, with high scores in all key metrics. CatBoost's ability to handle categorical features and gradient boosting contributed to its consistent performance. Logistic Regression, with its high scores across the board, also demonstrated its suitability for this dataset. MLP's performance could be attributed to its ability to handle complex and nonlinear data. Random Forest Classifier (RFC) yielded respectable scores, though slightly lower than the top-performing algorithms. Its performance may be influenced by its sensitivity to noise or randomness in the data. Naive Bayes, while still providing decent results, showcased its limitations in this dataset compared to other algorithms.

The ensemble methods—Bagging and Stacking—once again stood out for their impressive performances. The Bagging Ensemble method achieved high scores, suggesting its effectiveness in combining multiple models for stress prediction. The Stacking Ensemble method excelled with near-perfect scores, showcasing its ability to further improve predictions by intelligently combining various models' outputs.

4.3.3 Comparing the Datasets and Preferred Approaches

Comparing the results from the two datasets highlights some interesting insights. Despite

differences in their nature (real-world vs. simulated), both datasets consistently demonstrated the strengths of certain algorithms and methods.

CatBoost and Logistic Regression consistently performed well in both datasets, indicating their robustness across different data scenarios. The ability of CatBoost to handle categorical features and the generalizability of Logistic Regression contribute to their reliability. Ensemble methods, particularly Stacking, consistently produced remarkable results, suggesting that combining models can significantly enhance predictive accuracy. However, the potential risk of overfitting must be carefully considered, especially with the perfect scores obtained. Ultimately, the preferred approach might depend on the application's context and goals. If robustness and reliability are the primary concerns, CatBoost, Logistic Regression, or MLP might be favored. If the utmost accuracy is crucial, then the Stacking Ensemble method might be preferred, with awareness of potential overfitting challenges.

The comparison of results between the SWELL-KW Dataset and the Wearable Device Simulated Dataset underscores the importance of evaluating algorithms in diverse scenarios to ensure their adaptability and reliability across different contexts.

CHAPTER 5

CONCLUSION

This study contributes to the understanding of stress prediction using wearable devices and wireless body sensor networks. The results from both datasets highlight the strengths and limitations of various machine learning algorithms and ensemble methods in different scenarios. The preferred approach may vary based on the application's context and objectives. The findings support the central hypothesis that ensemble learning methods can enhance the accuracy and reliability of stress prediction. Both CatBoost and Logistic Regression consistently performed well across datasets, making them reliable choices for stress prediction applications. The Multi-Layer Perceptron demonstrated its robustness in handling complex data and emerged as a strong contender. However, it's important to note that while these algorithms displayed remarkable results, concerns about overfitting and generalizability must be taken into account.

Ensemble methods, specifically Bagging and Stacking, stood out for their ability to improve predictive accuracy by leveraging the strengths of multiple models. The Bagging Ensemble method demonstrated reliability in producing accurate predictions, while the Stacking Ensemble method pushed boundaries by achieving near-perfect scores. Nevertheless, the potential for overfitting in ensemble methods necessitates caution and the need for validation on real-world data.

Overall, this study underscores the importance of thoroughly evaluating machine learning algorithms and ensemble methods in various contexts to ensure their effectiveness and adaptability. Future research should focus on validating these findings with real-world data and exploring ways to mitigate potential overfitting issues in ensemble methods. As the field of wearable technology continues to evolve, the insights from this study can contribute to the development of more accurate and reliable stress prediction models for improved well-being and health management.

5.1 Future work

In light of the results and findings of this investigation, several pertinent avenues for subsequent research emerge:

Our analysis has provided insights on the effectiveness of the proposed algorithms using the SWELL-KW dataset, a representative of real-world data. To expand our understanding, it would be required to subject these algorithms to diverse real-world datasets. This approach is not merely a formality; it is central to evaluating the consistency and adaptability of the algorithms, ensuring their relevance and accuracy across a spectrum of real-life scenarios.

While the preliminary outcomes are promising, there remains an underlying concern of overfitting, particularly evident in the near-perfect scores produced by the Stacking Ensemble method. It is imperative to ascertain the generalizability of this model, ensuring that its performance isn't disproportionately tailored to our dataset. To that end, the adoption and experimentation with methodologies that can counteract overfitting will be pivotal.

Furthermore, the simulated data derived from wearable devices, inclusive of contextual parameters such as activity and ambient conditions, warrants a more in-depth exploration. Despite the comprehensive nature of the data, it is conceivable that certain facets, when appropriately harnessed, could further refine the predictive capabilities of our models. Identifying these salient features could be instrumental in elevating the efficacy of stress prediction.

While the current study provides a robust foundation, the horizon of research in this domain is expansive. Future endeavours should pivot around validating algorithmic consistency, safeguarding against overfitting, and unearthing nuanced data dimensions to augment stress prediction accuracy.

References

Banerjee, J. S., Mahmud, M., & Brown, D. (2023). Heart Rate Variability-Based Mental Stress Detection: An Explainable Machine Learning Approach. *SN Computer Science*, *4*(2), 176. https://doi.org/10.1007/s42979-022-01605-z

Bhushan, U., & Maji, S. (2023). *Prediction and Analysis of Stress Using Machine Learning: A Review* (pp. 419–432). https://doi.org/10.1007/978-981-19-3148-2_35

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1023/A:1018054314350

Chan, S. F., & La Greca, A. M. (2013). Perceived Stress Scale (PSS). In *Encyclopedia of Behavioral Medicine* (pp. 1454–1455). Springer New York. https://doi.org/10.1007/978-1-4419-1005-9_773

Cohen, S., Kamarck, T., & Mermelstein, R. (1983). Perceived Stress Scale. *APA PsycTests.* https://psycnet.apa.org/doiLanding?doi=10.1037%2Ft02889-000

Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Kluwer Academic Publishers, Boston.*

Dorogush, A. V., Ershov, V., & Gulin, A. (2018). *CatBoost: gradient boosting with categorical features support*. http://arxiv.org/abs/1810.11363

Epel, E. S., Crosswell, A. D., Mayer, S. E., Prather, A. A., Slavich, G. M., Puterman, E., & Mendes, W. B. (2018). More than a feeling: A unified view of stress measurement for population science. *Frontiers in Neuroendocrinology*, *49*(December 2017), 146–169. https://doi.org/10.1016/j.yfrne.2018.03.001

Gjoreski, M., Gjoreski, H., Luštrek, M., & Gams, M. (2016). How Accurately Can Your Wrist Device Recognize Daily Activities and Detect Falls? *Sensors*, *16*(6), 800. https://doi.org/10.3390/s16060800

HSE. (2022). HSE publishes annual work-related ill-health and injury statistics for 2021/22. In *British Health and Safety Executive (HSE)*. https://press.hse.gov.uk/2022/11/23/hse-publishes-annual-work-related-ill-health-and-injury-statistics-for-2021-22/

Kobsar, D., Charlton, J. M., Tse, C. T. F., Esculier, J.-F., Graffos, A., Krowchuk, N. M., Thatcher, D., & Hunt, M. A. (2020). Validity and reliability of wearable inertial sensors in healthy adult walking: a systematic review and meta-analysis. *Journal of NeuroEngineering and Rehabilitation*, *17*(1), 62. https://doi.org/10.1186/s12984-020-00685-3

Koldijk, S., Neerincx, M. A., & Kraaij, W. (2018). Detecting Work Stress in Offices by Combining Unobtrusive Sensors. *IEEE Transactions on Affective Computing*, *9*(2), 227–239. https://doi.org/10.1109/TAFFC.2016.2610975

Li, X., Dunn, J., Salins, D., Zhou, G., Zhou, W., Schüssler-Fiorenza Rose, S. M., Perelman, D.,

Colbert, E., Runge, R., Rego, S., Sonecha, R., Datta, S., McLaughlin, T., & Snyder, M. P. (2017). Digital Health: Tracking Physiomes and Activity Using Wearable Biosensors Reveals Useful Health-Related Information. *PLOS Biology*, *15*(1), e2001402. https://doi.org/10.1371/journal.pbio.2001402

Lin, K., Jie, B., Dong, P., Ding, X., Bian, W., & Liu, M. (2022). Convolutional Recurrent Neural Network for Dynamic Functional MRI Analysis and Brain Disease Identification. *Frontiers in Neuroscience*, *16*. https://doi.org/10.3389/fnins.2022.933660

McDonald, M. M., Khoo, W. H., Ng, P. Y., Xiao, Y., Zamerli, J., Thatcher, P., Kyaw, W., Pathmanandavel, K., Grootveld, A. K., Moran, I., Butt, D., Nguyen, A., Warren, S., Biro, M., Butterfield, N. C., Guilfoyle, S. E., Komla-Ebri, D., Dack, M. R. G., Dewhurst, H. F., … Phan, T. G. (2021). Osteoclasts recycle via osteomorphs during RANKL-stimulated bone resorption. *Cell*, *184*(5), 1330-1347.e13. https://doi.org/10.1016/j.cell.2021.02.002

Memon, A., Taylor, K., Mohebati, L. M., Sundin, J., Cooper, M., Scanlon, T., & De Visser, R. (2016). Perceived barriers to accessing mental health services among black and minority ethnic (BME) communities: A qualitative study in Southeast England. *BMJ Open*, *6*(11), 1–9. https://doi.org/10.1136/bmjopen-2016-012337

Pabreja, K., Singh, A., Singh, R., Agnihotri, R., Kaushik, S., & Malhotra, T. (2021). *Stress Prediction Model Using Machine Learning* (pp. 57–68). https://doi.org/10.1007/978-981-15-4992-2_6

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408. https://doi.org/10.1037/h0042519

Schmidt, R. M. (2019). *Recurrent Neural Networks (RNNs): A gentle Introduction and Overview*. *1*, 1–16. http://arxiv.org/abs/1912.05911

Tawakol, A., Ishai, A., Takx, R. A., Figueroa, A. L., Ali, A., Kaiser, Y., Truong, Q. A., Solomon, C. J., Calcagno, C., Mani, V., Tang, C. Y., Mulder, W. J., Murrough, J. W., Hoffmann, U., Nahrendorf, M., Shin, L. M., Fayad, Z. A., & Pitman, R. K. (2017). Relation between resting amygdalar activity and cardiovascular events: a longitudinal and cohort study. *The Lancet*, *389*(10071), 834–845. https://doi.org/10.1016/S0140-6736(16)31714-7

Wang, D., & Yue, X. (2019). The Weighted Multiple Meta-Models Stacking Method for Regression Problem. *2019 Chinese Control Conference (CCC)*, 7511–7516. https://doi.org/10.23919/ChiCC.2019.8865869

Zaccaro, A., Piarulli, A., Laurino, M., Garbella, E., Menicucci, D., Neri, B., & Gemignani, A. (2018). How Breath-Control Can Change Your Life: A Systematic Review on Psycho-Physiological Correlates of Slow Breathing. *Frontiers in Human Neuroscience*, *12*.

https://doi.org/10.3389/fnhum.2018.00353

Zhang, J., Yin, H., Zhang, J., Yang, G., Qin, J., & He, L. (2022). Real-time mental stress detection using multimodality expressions with a deep learning framework. *Frontiers in Neuroscience*, *16*. https://doi.org/10.3389/fnins.2022.947168

Zhang, S., Li, Y., Zhang, S., Shahabi, F., Xia, S., Deng, Y., & Alshurafa, N. (2022). Deep Learning in Human Activity Recognition with Wearable Sensors: A Review on Advances. *Sensors*, *22*(4), 1476. https://doi.org/10.3390/s22041476