Predicting Stress Levels from Multimodal Wearable Sensor Data

Omar Moustafa (900222400)

Nour Kahky (900221042)

Omar Abdelazim (900222793)

DSCI 4411

Fall 2025

Introduction

Wearable devices have become increasingly capable of tracking physiological responses on a continuous basis. This recent technological advancement provides the opportunity to monitor real-time stress levels, potentially allowing early detection and intervention, necessary mental health support, and personalized wellbeing applications. Stress is well-known to lead to severe changes across multiple different physiological systems, such as the autonomic nervous system and cardiovascular activity. Specifically, physiological stress activates the sympathetic nervous system which leads to measurable physical changes such as increased sweat gland activity, elevated heart rate, and reduced Heart Rate Variability (**HRV**), as represented by the flowchart below.

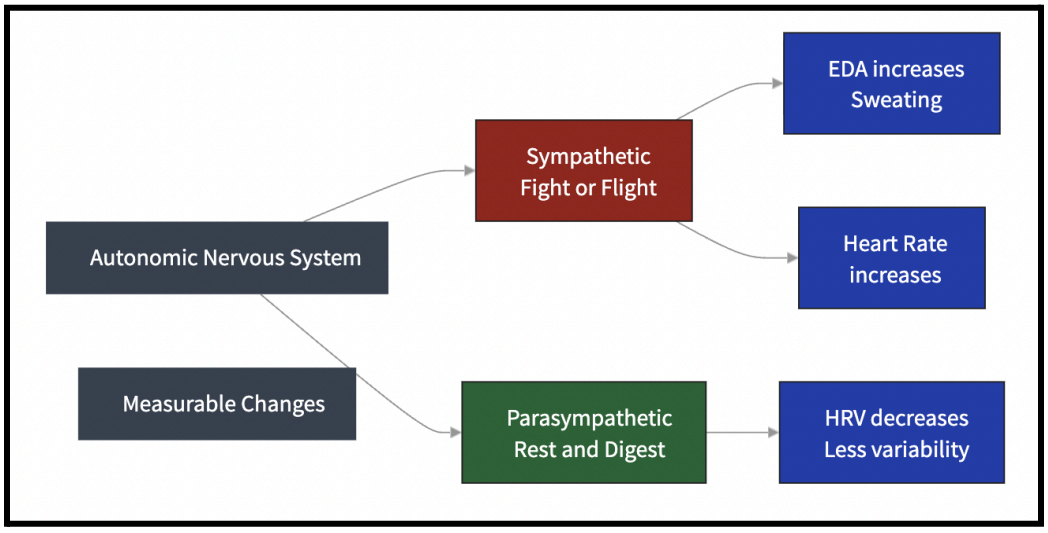

Figure 1

These particular physical changes can be measured using modern wearable technology. The specific piece of wearable technology utilized in this study is the **Empatica E4 Wristband**, which is a functioning smartwatch that includes several sensors that are relevant to stress detection and analysis, as summarized by the following table.

Signal:	**Sampling Rate:**	**What It Measures:**
EDA	4 Hz	Skin conductance
HR	1 Hz	Beats per minute
ACC	32 Hz	3-axis movement
BVP	64 Hz	Pulse waveform
TEMP	4 Hz	Skin temperature
IBI	Variable	Beat-to-Beat intervals

Table 1

Prior studies and related work have demonstrated that EDA and HR are strong indicators of stress, while exercise complicates stress detection due to overlapping physiological signatures. Traditional ML models, namely the Random Forest (**RF**) model, and neural networks (**NNs)** were previously used, with classification having a tendency to outperform regression models. Another possible option is real-time processing, however this is rarely implemented but is conceptually based on sliding window prediction.

Building on this foundation, the main objective of this project is to build and evaluate data mining models capable of identifying stress episodes in real-time from wearable-sensed data. The study focuses on three predictive models and tasks: (1) **classifying whether a user is stressed, resting, or engaged in physical activity**, (2) **predicting continuous stress levels based on self-reported ratings**, and (3) **anomaly detection.** By implementing cohesive machine learning (**ML**) models, insightful data mining techniques, and then thoroughly analyzing their performances on these tasks, this project sets out to provide valuable insights into both the strengths and limitations of stress detection using wearable sensors and stress-detecting devices.

Dataset Description

The dataset includes physiological recordings from 100 participants across three different controlled environments: **Rest (Baseline)**, **Stress (TSST)**, and **Physical Activity (Exercise).** This structured contrast provides grounds for a comprehensive basis in evaluating multiple stress-detection models, as the conditions produce separable physiological signatures that are suitable for supervised learning. Additionally, continuous subjective stress ratings were collected throughout the study to facilitate regression analysis as well.

All of the participants were subjected to completing each condition in either a fixed or semi-randomized order. The cohort was further divided into three groups (**STRESS**, **AEROBIC**, and **ANAREOBIC**) based on the primary activity that was performed. The distribution of participants across these groups is shown in the figure below.

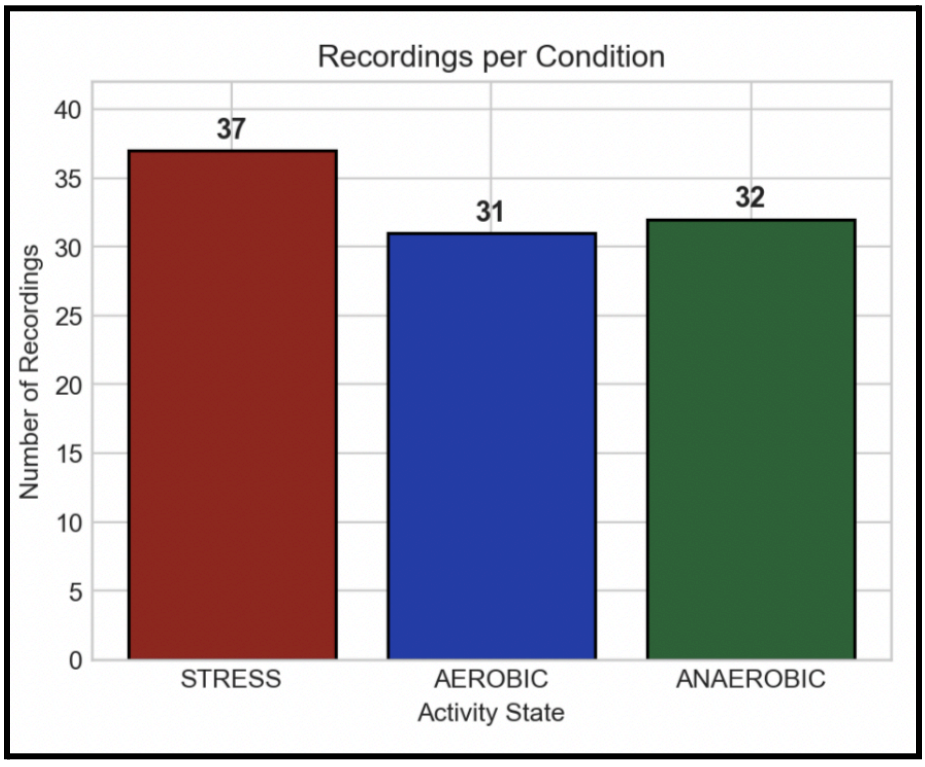

Figure 2

To further analyze the data, the raw data was analyzed to display example signals for one participant during the stress condition. To point out, the x-axis for all three figures is time, recorded in seconds, going on until 2224 total seconds. Looking deeper into the following figure, starting with the top panel, this represents the EDA, which is the skin conductance signal. It can be seen that it varies over time as the participant goes through different protocol stages. Moving onto the middle panel, that is the heart rate of the participant. Again, there is some variation but too nothing dramatic since they are just sitting down. As for the third panel at the very bottom, this represents the accelerometer. It can be seen that it is basically flat, this is also because the person is seated during the stress test and not moving around. Therefore, for this particular panel, it can be inferred that it would look very different if the participant was exercising.

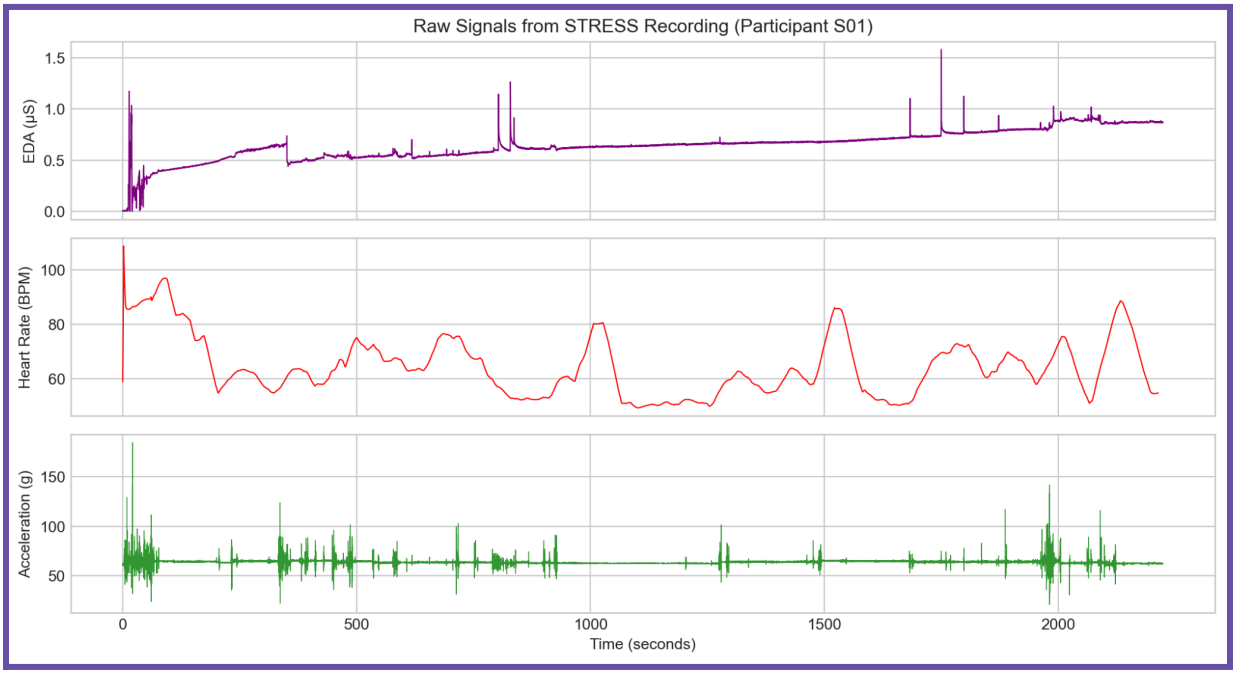

Figure 3

Methodology

Feature Extraction:

Following the data preprocessing stage, Feature Extraction was performed which can be summarized using the following flowchart.

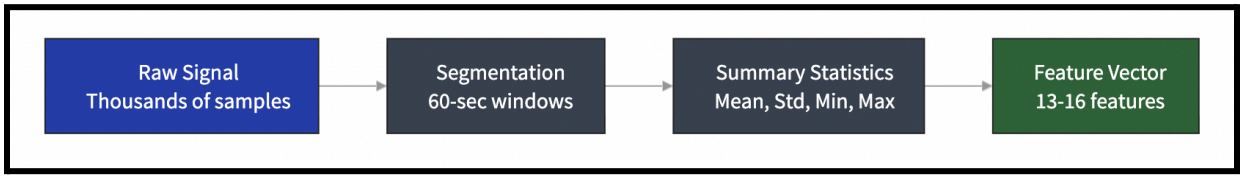

Figure 4

The original dataset contained thousands of raw samples which were segmented into either 60-second windows or based on tags in the data that mark protocol stages. For each segment, summary statistics were computed and the data was left with a feature vector of 13 to 16 numbers per segment. The following list contains everything that was computed:

- ***EDA:*** mean, std, min, max, range (5 features)
- ***HR:*** mean, std, min, max, range (5 features)
- ***ACC:*** mean, std (2 features)
- ***Duration:*** duration of the segment (1 feature)
- ***For stress prediction:*** IBI mean, std, RMSSD (3 more)

After running feature extraction on all recordings, the segments that were either too long or too short were filtered out and outliers were clipped to reasonable physiological ranges. After these early-stage processes, there was a remaining total of several-hundered segments where the breakdown of them by state is shown below. It can be interpreted that STRESS has the most segments because the protocol has multiple stages that each become a segment.

After feature extraction:

- 771 segments total
- 100 recordings processed
- Segments filtered to 30-600 seconds
- Outliers clipped to plausible ranges

Segments by state:

State	Segments
STRESS	290
AEROBIC	257
ANAEROBIC	224

Figure 5

To conclude the feature extraction, the following figure represents a group of boxplots which show the distributions of each feature. Starting with the top-left box plot, which is of the heart rate mean, it can be clearly seen that STRESS is lower and tighter, while AEROBIC and ANAEROBIC are higher with more spread. For the next boxplot, heart rate standard deviation, this shows ANAEROBIC being clearly the highest one of the three. This observation does make scientific sense because interval training contains periods of both high and low intensity. From the EDA boxplots overall, these tend to show more overlap between the three groups and not as clean of separations. The main takeaway from this group of boxplots is that some features separate the classes well, whereas some do not.

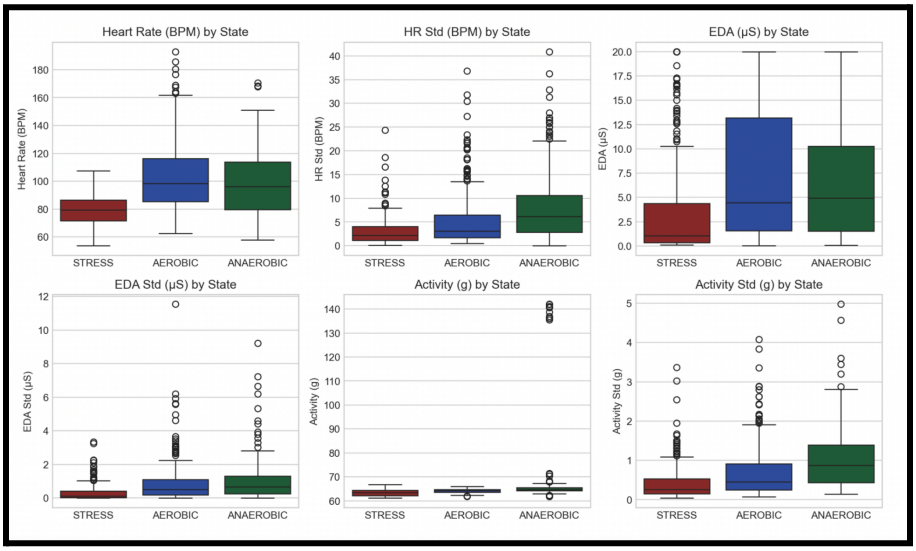

Figure 6

Tasks & Results

Task 1: Activity State Classification

The goal of this first task is to accurately classify segments into **STRESS**, **AEROBIC**, or **ANAEROBIC**. This was fulfilled by building an RF classification model which consisted of 300 trees and balanced weights. The setup of the model was to split the dataset into 75% to be the training set and the remaining 25% to be the testing set, with the training set normalized using the **StandardScaler()** function in Python. The overall **accuracy** was **82.9%** and the resulting confusion matrix can be seen directly below.

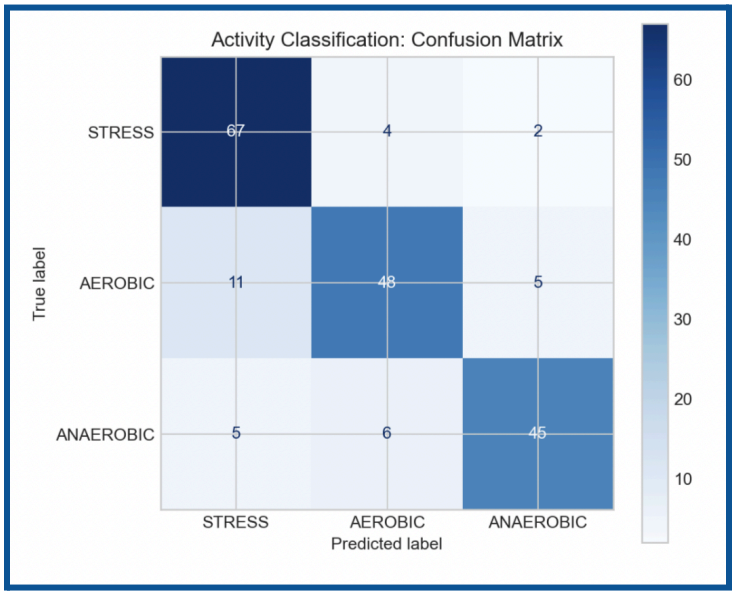

Figure 7

From this confusion matrix, it can be seen that STRESS is very easy to classify with 67 correct classifications, which are known as True Positives (**TPs**), and only 6 misclassifications (4 were misclassified as AEROBIC and 2 were misclassified as ANAEROBIC). This task was further evaluated with a full and cohesive classification report as revealed and highlighted by the following table.

Class	Precision	Recall	F1-Score	Support
AEROBIC	0.83	0.75	0.79	64
ANAEROBIC	0.87	0.80	0.83	56
STRESS	0.81	0.92	0.86	73
Weighted Avg	**0.83**	**0.83**	**0.83**	193

Table 2

To evaluate the results of this task, the evaluation metrics consisted of those that validate the regression accuracy (**RMSE** and **MAE**) as well as those that validate the ranking quality (**Accuracy**, **Precision**, **Recall**, **F1**, and **confusion matrices**). Therefore, building off the success of very accurately classifying STRESS segments as such, this is supported by having a Recall value of 0.92, further emphasizing that it rarely ever gets missed. Thus, this confirms that the model can reliably tell psychological stress apart from physical exercise.

To know which features were most important for this classification task thus far, the following **Feature Importance** plot reveals that the three most important features were **duration_s**, **acc_mean**, **hr_max**, and **hr_mean**. Looking deeper into this top four, duration being the most important is partly due to how different conditions have different typical segment lengths. The second most important feature was accelerometer mean. This makes intuitive sense since, as shown in Figure 3, where the accelerometer signal was almost flat during the stress test, individuals are moving much more during exercise conditions (AEROBIC and ANAEROBIC) than during the seated STRESS condition. The two subsequent most important features are both heart rate features, namely hr_mean, hr_max, which does make sense given what was observed in the group of boxplots in Figure 6. Additionally, it can be observed below that EDA features were among the less important features for this task. Consequently, movement and heart rate

patterns demonstrated more discrimination for differentiating between these three activity states, even though EDA is frequently mentioned in the literature as the main stress indication.

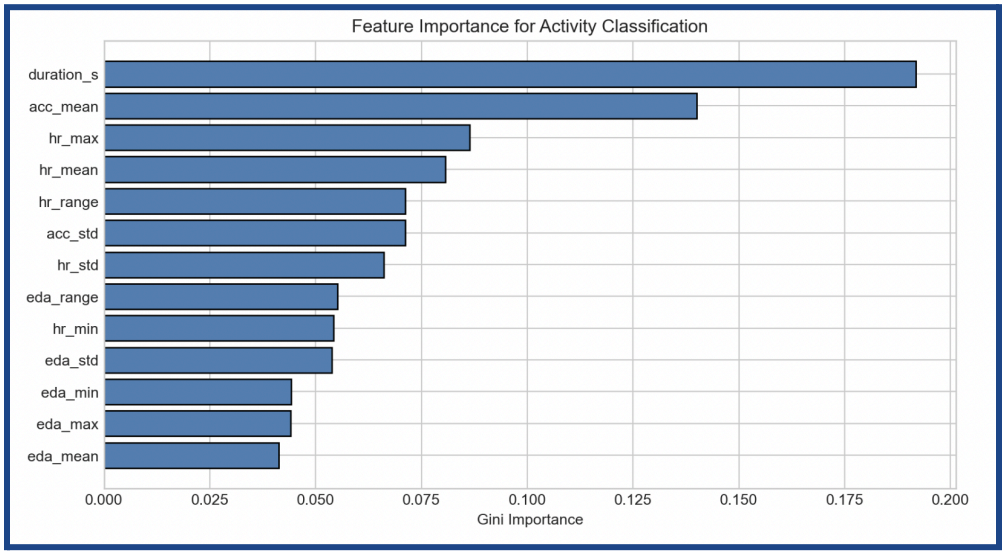

Figure 8

Next was a model comparison of three different tasks performing this task, which were Random Forest, as previously mentioned, and the two others being AdaBoost and Gradient Boosting. Their comparisons can be seen in the following figure.

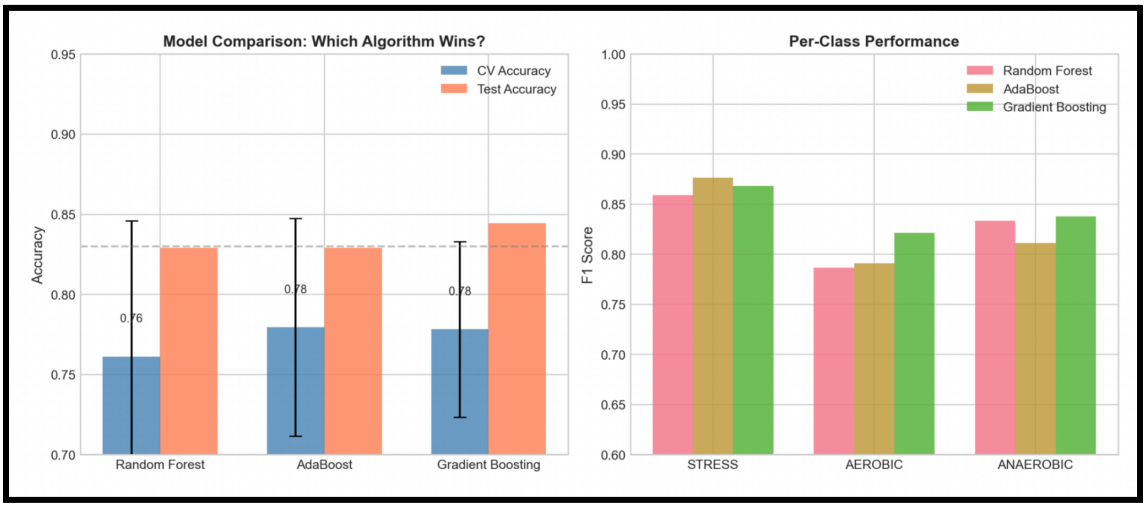

Figure 9

Figure 10

Starting with Figure 9 on the left, all three classification algorithms, Random Forest, AdaBoost, and Gradient Boosting, achieved a similar test accuracy with both Random Forest and AdaBoost achieving approximately 83% and Gradient Boosting achieving closer to 85%. This reveals that all three models perform similarly, specifically within about 2% accuracy of each other. The cross-validation (**CV**) variance displayed by the error bars is comparatively small for all three models, indicating robust and consistent performance across various data splits. This, therefore, suggests that the results are indeed reliable and not significantly influenced by the way the data was divided.

As for Figure 10 on the right, looking at per-class F1 scores, all three models do well on STRESS as it is consistently well-classified across all models. The main differences are in how they handle AEROBIC and ANAEROBIC, with AEROBIC showing the most variation between algorithms. The models tend to agree most of the time, but disagreements happen in about 13% of cases, mostly in ambiguous regions between exercise types.

Unsupervised clustering was also run to see how data naturally groups. Using **K-Means Clustering** with 2 clusters, the data splits mainly by heart rate and clusters naturally align with activity states. Looking at state composition, the low-HR cluster is mostly STRESS samples, while the high-HR cluster has exercise states. This, therefore, confirms that the features do indeed capture meaningful physiological differences and validates the feature engineering approach done in the very beginning.

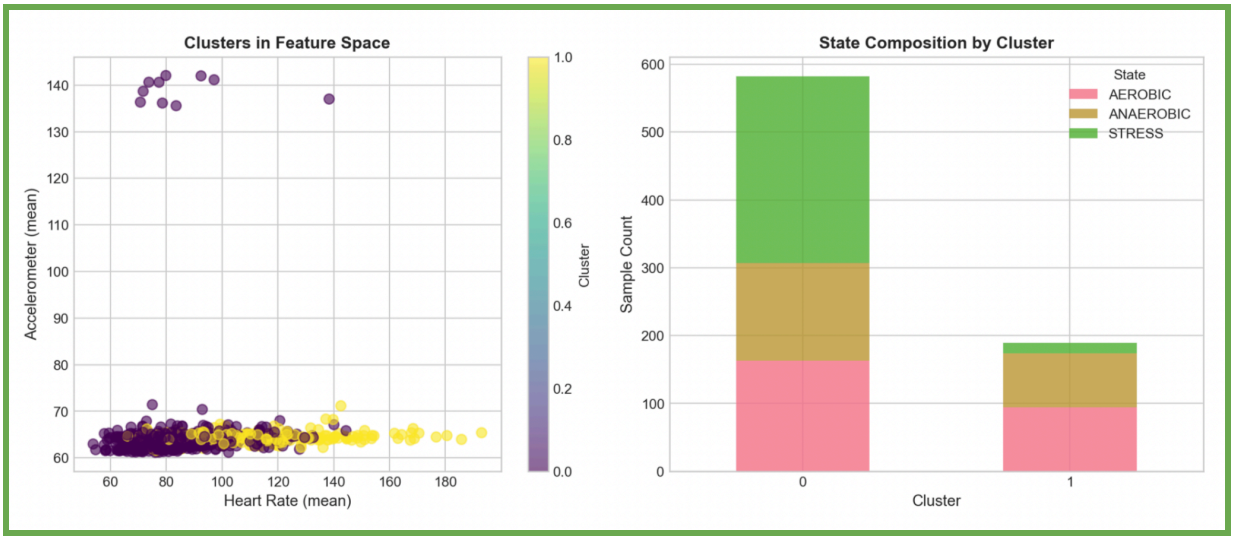

Figure 11

Figure 12

Task 2: Stress Level Prediction

The second task set out to predict subjective stress rating (**0-10**) from physiological features. One standout reason why this is a harder task than the previous one is because there were only 108 ratings from 18 participants, leading to an ever smaller sample size to work with. More important reasons why this is a harder task are that self-reported stress levels can be noisy, physiological changes during psychological stress are subtle, and people respond to stress differently.

The first step of this task was to conduct trier-style stress protocol with multiple stages, which worked as expected with the Trier Mental Challenge Test (**TMCT**) having the highest stress levels (**mean = 6.4**) and rest periods having the lowest stress levels. Both the distribution of self-reported stress and the average stress levels of each protocol stage can be observed in the following figures.

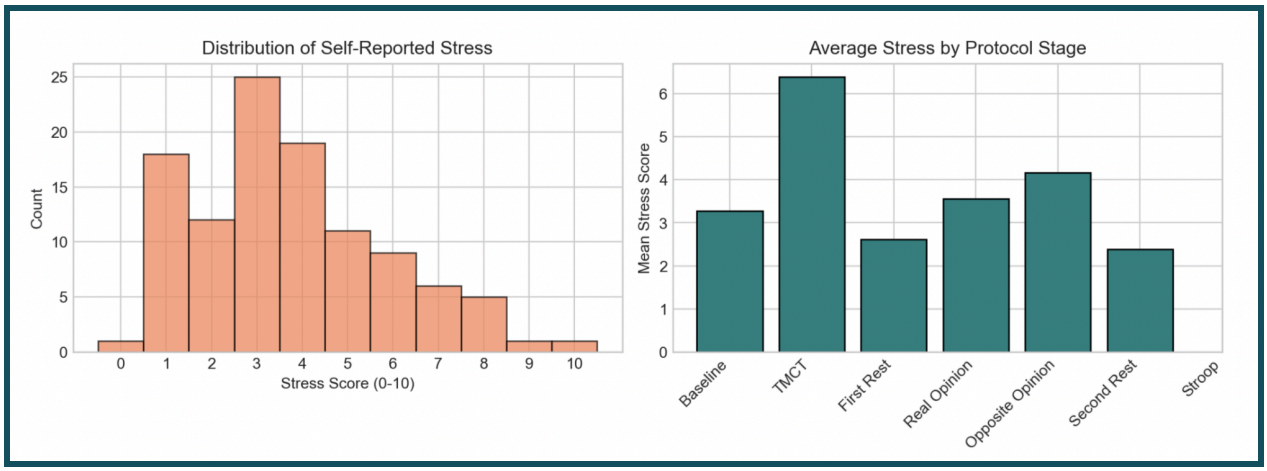

Figure 13

Figure 14

The distribution of self-reported stress shows that most of the values are in the 2-6 range with few extreme values. As for protocol, it contains multiple subsequent stages. The stages begin with Baseline, which is just sitting calmly. TMCT is a mental arithmetic task. Then some rest. Then opinion tasks where the participant will have to argue positions. Then some more rest. Then the test ends with Stroop, which is the classic color-word interference task. It can be seen that the opinion tasks are the ones that produce the highest stress on average whereas rest periods generate much lower stress. So the protocol is working as it is inducing various stress levels depending on the various stages.

As hinted by the description of the performed tasks, this was not a classification task, but rather a regression task. Using the same model as before, Random Forest, but now for regression, **5-fold cross-validation** (**CV**) was computed. Looking at the figure on the left, the dashed line is $\mathbf{R^2 = 0}$, which is what one would get by just predicting the mean every time. Specifically looking at the bars, some of them are actually negative, indicating worse outcomes than the baseline. As for the scatter plot on the right, if the predictions were correct, points would line up along the diagonal. However, this is not the case as more of a blob is shown and the R^2 is very close to zero. These poor results demonstrate the lack of predictive power of the model, indicating that

this second task was a failed task. This regression analysis is based on the two figures seen directly below.

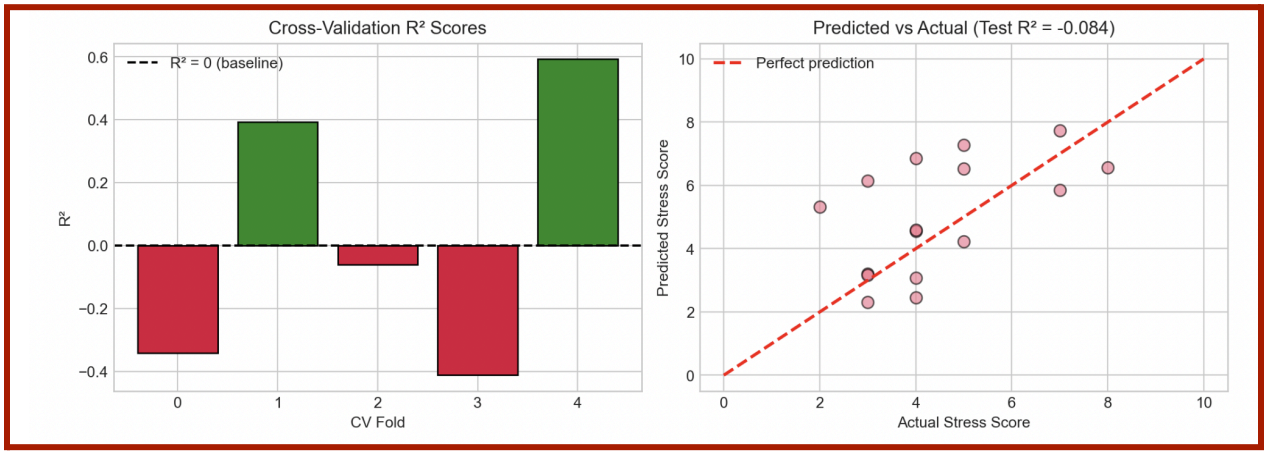

Figure 15

Figure 16

Task 3: Anomaly Detection

The third and final task conducted in this project was an anomaly detection-based task. This task consisted of training on baseline rest data and seeing if stressful periods happened to look unusual or not. In the figure below, Figure 17, the blue bars represent what fraction of baseline samples get flagged as anomalies (which is expected to be low). On the other hand, the orange bars represent the more stressful samples. Out of the three implemented models for this task, **One-Class SVM** performed the best as it flagged approximately 46% of stressful periods as anomalies, compared to only 18% of baseline, which is a **2.6x** difference. This can be described as a modest success as it shows its partial capabilities at detecting stress as deviations from baseline, even if it can not predict the exact stress levels.

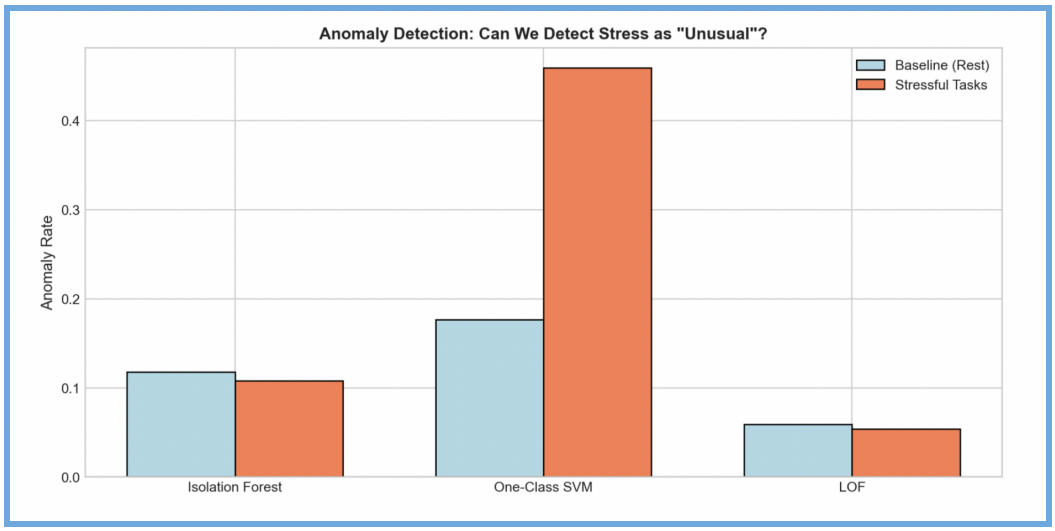

Figure 17

Going into more of a numeric comparison of these three implemented models, the following table reveals that, as previously mentioned, One-Class SVM flagged 2.6x more anomalies flagged during stressful tasks, that is almost triple the times of anomalies flagged flagged by Isolation Forest and LOF, both at only **0.9x**. This major distinction indicates that, in comparison to the other approaches, One-Class SVM more accurately depicts typical resting physiology. Stress and baseline samples were identified at about identical rates by Isolation Forest and LOF, both of which displayed anomaly ratios near 1.0, revealing poor differentiation between the two states.

Model	Baseline Flagged	Stressful Flagged	Ratio
Isolation Forest	11.8%	10.8%	0.9x
One-Class SVM	17.6%	45.9%	2.6x
LOF	5.9%	5.4%	0.9x

Table 3

Summary of the Tasks

Task	N	Method	Performance	Outcome
Activity Classification	771	Random Forest	83% accuracy	Worked
Stress Regression	63	Random Forest	$R^2 = 0.03$	Failed
Anomaly Detection	71	One-Class SVM	2.6x detection ratio	Partial

Table 4

Now to provide an overarching summary of the three conducted tasks. Starting with Activity State Classification, this task worked well as it achieved over 80% accuracy. This first task confirmed that working models can reliably distinguish between participants that are resting and that are exercising. The second task, Stress Level Prediction using regression, ended up failing with a negative R^2 that was near zero as well. As for Anomaly Detection at the very end, this showed partial success where detecting stress periods that are unusual compared to the given baseline can be done to a decent degree, but not perfectly. Overall, to tell apart an exercising participant from a sitting participant is straightforward, predicting stress intensity is difficult, and detecting outlier levels of induced stress shows some promising results that can be improved to reach greater levels of success.

Key Takeaways

Looking into the valuable insights that were obtained throughout this project, there were five key takeaways that stood out:

1. **Wearable devices can accurately distinguish between physiological states** with approximately 83% accuracy for activity classification. Specifically, modern wearable devices work for classifying activity states, namely differentiating between stress and exercise.
2. **Predicting stress intensity does not work** well **with** the provided **simple features**.
3. **Sample size matters** where in this case the number of contributing participants were not enough to capture variability in stress responses.
4. **Negative results are useful**, knowing the limits prevents overpromising
5. Consumer wearable stress claims should be viewed with skepticism

Future Work

After evaluating and discussing the results of the conducting data mining tasks, it should be noted what can be done in the future to improve this. For one thing, instead of computing summary statistics, implementing and using time series analysis and models such as Long Short-Term Memory (**LSTM**) or Convolutional Neural Networks (**CNNs**) that can capture complex temporal patterns. Another potential improvement is that rather than using one model for all of the participants, personalized baselines for each person could be used. Additionally, instead of just lab-data, collecting real-world ambulatory data could be of real use. Collectively,

these would all be beneficial adjustments and good directions for future work in this particular field of research.

Conclusion

In conclusion, this project thoroughly demonstrated that wearable sensor data can effectively classify discrete activity states but comes across major obstacles when trying to predict subjective stress intensity. The Random Forest classification achieved 82.9% accuracy in distinguishing between STRESS, AEROBIC, and ANAEROBIC conditions, with particularly strong performance on psychological stress detection, as reported by its 92% recall value. However, regression-based prediction of continuous stress ratings were deemed unsuccessful, highlighting the complexity of translating into ongoing subjective circumstances. Finally, the anomaly detection approach showed decent potential, with One-Class SVM successfully identifying 46% of periods of stress as anomalies. Overall, these findings underline both the potential and current limitations of consumer wearable devices for stress tracking, heavily emphasizing the necessity for more sophisticated modeling approaches as well as personalized baselines in future work and research.

References

Boucsein, W. (2012). *Electrodermal activity* (2nd ed.). Springer. https://doi.org/10.1007/978-1-4614-1126-0

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Empatica. (2023). *E4 wristband technical specifications*. https://www.empatica.com/research/e4/

Hongn, A., Bosch, F., Prado, L. E., Ferrández, J. M., & Bonomini, M. P. (2025). Wearable physiological signals under acute stress and exercise conditions. *Scientific Data*, *12*(1). https://doi.org/10.1038/s41597-025-04845-9

Hongn, A., Bosch, F., Prado, L., & Bonomini, P. (2025). *Wearable device dataset from induced stress and structured exercise sessions* (Version 1.0.1). PhysioNet. https://doi.org/10.13026/he0v-tf17

Kirschbaum, C., Pirke, K.-M., & Hellhammer, D. H. (1993). The 'Trier Social Stress Test' – a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology*, *28*(1-2), 76–81. https://doi.org/10.1159/000119004

Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. (1996). Heart rate variability: Standards of measurement, physiological interpretation, and clinical use. *Circulation*, *93*(5), 1043–1065. https://doi.org/10.1161/01.CIR.93.5.1043