Individual Diploma Thesis

ANALYSIS OF REAL-TIME E4-BASED PSYCHOPHYSIOLOGICAL DATA FOR MACHINE-LEARNING BASED CLASSIFICATION

Sotiris Zenios

University of Cyprus

Department of Computer Science

May 2023

University of Cyprus

Department of Computer Science

Analysis of Real-time E4-based Psychophysiological Data and Machine-learning-based Classification

Sotiris Zenios

Supervisor

Chryssis Georgiou

Diploma project has been submitted for partial fulfillment of the requierements of Informatics Degree acquisition from the University of Cyprus

May 2023

Acknowledgement

I am incredibly thankful to my supervisor, Dr. Chryssis Georgiou, Professor at the Department of Computer Science at the University of Cyprus, for giving me the opportunity to work on this topic and for the assistance he provided me throughout my thesis.

I am also immensely grateful to the Department of Psychology at the University of Cyprus for providing us with the experimental data samples.

Additionally, I want to show my appreciation to Andria Trigeorgi for providing me with pertinent information related to the topic and for their continued support and feedback.

Lastly, I am very thankful to my parents and friends for the moral support they gave me. Without them, this study would not be possible.

Abstract

In recent years, there has been a swift advancement in both wearable technologies and Artificial Intelligence. This thesis seeks to analyze the potential of these two working together to detect and prevent dysfunctional pain coping. Specifically, this thesis will investigate whether wearable devices can accurately distinguish between functional and dysfunctional pain coping strategies during real life situations.

In an experiment conducted by the Department of Psychology of the University of Cyprus, all participants were instructed on how to wear a wearable device, the Empatica E4, and were prompted to respond to questions on an app pre-installed on provided smartphones. Specifically, participants answered questions about social context, experiences of stress or pain (both physical and emotional), and the use of coping strategies. During the experiment, a variety of psychophysiological signals, Photoplethysmography (PPG), Electrodermal Activity (EDA), Accelerometer (ACC) and Temperature (TEMP) — were recorded in real time. After obtaining the raw signals from the specific time windows that participants answered questions, psychophysiological features were extracted. Using two different feature selection methods, the most significant features were chosen. Using those features, a number of supervised Machine Learning models (Adaptive Boosting, Gradient Boosting Decision Tree, Random Forest, and Extra Trees) were employed in order to find the best and most effective one.

This analysis shows that the most important features, in order to classify people into the two categories, are features derived from heart rate variability.The Gradient Boosting Decision tree, across different scenarios, when fine-tuned can correctly classify people, with its accuracy reaching 70% in the general scenario.

It was also shown that people who considered themselves to not belong in any of the two aforementioned categories, acceptance or avoidance, were closer categorised into the avoidance class. Moreover, data acquired by the watch can give results with performance similar to those acquired by the stationary device.

Contents

Chapter 1

Introduction

Contents

1.1 Motivation

The last decade has seen a significant rise in the use of wearable technology, such as smartwatches and smart bands. These devices are able to track various psychophysiological signals, such as heart and sweat gland activity. These indicators, known as psychophysiological signals, have been demonstrated to be indicative of an individual's emotional response. Examples of such psychophysiological signals are Electrocardiogram (ECG) [1], Electrodermal Activity (EDA) [2], and facial electromyography (fEMG) [3].

A number of previous studies [4] [5] [6] [7] analysed such data. Firstly, previous works concentrated on signals recorded from stationary devices, with the exception of [6] which included measures from wearable devices. Moreover, the only features that were examined and used to train the models, are HRV time-domain features (measures used to quantify the amount of variation in the intervals between heartbeats over a specific period of time) [8] that come from ECG. The purpose of this study is to explore the subject in more depth so as to help health care.

1.2 Goals of the Study

This study makes use of data collected from an experiment regarding pain management techniques that was conducted by the Department of Psychology of the University of Cyprus. The ultimate goal of the present thesis is to contribute to the integration of a form of psychotherapy, called Acceptance and Commitment Therapy, in the everyday life. Acceptance and Commitment Therapy (ACT) [9] involves encouraging people to try to deal with their thoughts and feelings instead of blaming themselves about them or trying to ignore them. ACT is a vital therapy as it can help people struggling with OCD, anxiety, depression, etc. It separates people into two groups, based on their reactions at a certain time. The first group is 'acceptance', also known as 'functional', and it contains people who accept their problems and try to cope with them head-on. Contrary, the 'avoidance' group, which is also known as 'dysfunctional', is the exact opposite and includes people that deny to remain in contact with their thoughts and sensations and attempt to avoid them. An individual does not always fall into the same category; their classification changes depending on the environment and the circumstances. This thesis aims to effectively classify individuals into functional or dysfunctional regarding pain coping.

More specifically, focus is given on discovering whether signals recorded from wearable devices are sufficient to effectively train Machine Learning algorithms and to examine whether features extracted from signals, like PPG which will be explained in detail in Section 2.1.2, can be more effective than ECG. Moreover, this work is intended to examine feature selection methods from two different categories, in order to compare the results and conclude to the best-performing subsample of features.

1.3 Methodology

The methodology of the current study can be seen in Figure 1.1.

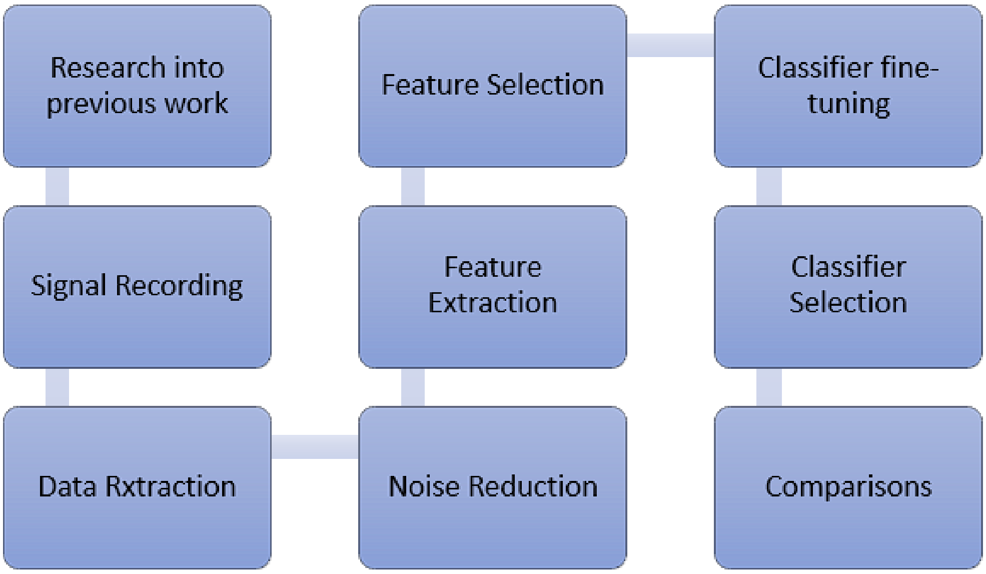


Figure 1.1: Methofology of current work

Firstly, the Department of Psychology of the University of Cyprus conducted an experiment in the lab, where psychophysiological signals were recorded from the Empatica E4 wearable device. These signals are Photoplethysmography (PPG), Electrodermal Activity (EDA), Accelerometer (ACC) and Temperature (TEMP).

Before analyzing the data that was gathered, a lot of attention was given to examining the efforts of the past years. For this reason, all the techniques used were evaluated, and the most effective ones were used in the current work.

Afterwards, with the use of Python, all signals were cleared from noised caused by the device's errors. Thus, the next step was to extract the necessary features from each one of the psychophysiological signals. Main focus was given to time-domain and statistical features.In order to train the algorithms quickly and efficiently, it was necessary to select only the most critical features. This was done using features extracted from two different methods, namely, Random Forest Classifier feature importance and SelectKBest algorithm, which will be explained in detail in Section 5.1.

Five distinct supervised binary classification Machine Learning algorithms were investigated, including Adaptive Boosting, Gradient Boosting Decision Tree, Bagging Decision Tree, Random Forest, and Extra Trees (Analysed in Section 2.2). These algorithms have parameters that require tuning, depending on the data, for optimal performance. Consequently, classifier fine-tuning was conducted, testing different parameter values for each algorithm and selecting the best ones. The subsequent step involved training the five algorithms using the chosen parameters from the previous step and selecting the best-performing one.

Finally, alternations of the original dataset were created, in order to better test different cases regarding the pain-coping strategy that the patients used. Those results were compared in order to better understand the correlation between the questionnaires and the recorded signals.Finally, comparison of results was a necessity.

1.4 Document Organization

The rest of this thesis is split into six chapters. Table 1.1 reports the content of each chapter

Chapter Number	Chapter Description
2	Overviews the background knowledge on which the thesis was built on. At first, the psychophysiological signals that were recorded in the Psychology lab are explained. Thereafter, the machine learning algorithms used are analysed, as well as the methodology and metrics used for evaluating the models. Finally, the devices that were used throughout the experiments are briefly described.
3	Explains in detail the four previous experiments.
4	Explains all the work that was done to obtain the features used to train the algorithm from the raw signals.It explaines the methodology used to get the samples from the patients 3-day recordings.Then, it explains the features that were extracted.
5	It analyses the feature selection methods and concludes to the selected subset of features for each dataset.
6	Describes the comparisons that were performed.
7	Summarizes the work done and suggests future improvements

Table 1.1: Document Organisation

Chapter 2

Background Knowledge

Contents

2.1 Psychophysiological Signals

The Department of Psychology of the University of Cyprus conducted a series of four experiments, which are explained in Section 3. In the experiment studied, the psychophysiological signals that are explained in this section were collected.

2.1.1 Electrocardiogram (ECG)

The heart generates electrical signals when it beats. These electrical signals can be noninvasively captured from the body's surface using an electrocardiogram (ECG) [10]. This electrical activity's basic pattern consists of three waves, referred to as P, QRS, and T [11] as seen in Figure 2.1. The ECG signal can extract three groups of features: frequency-domain, spectral, and time-domain. However, the focus of this study is primarily on time-domain features, as they are more applicable to our research, as suggested by a previous study [12]. Time-domain measures primarily concentrate on Heart Rate Variability (HRV), which refers to the fluctuations in the time intervals between successive heartbeats, specifically, the RR intervals. The RR intervals are the duration between two sequential R peaks in the ECG signal, with the R wave being part of the QRS complex.

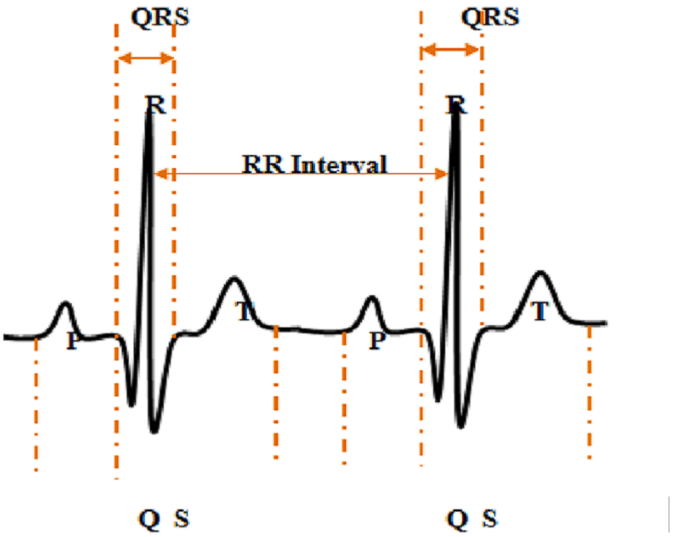


Figure 2.1: Visual Representation of ECG Signal

2.1.2 Photoplethysmography (PPG)

During a cardiac cycle, which spans from the start of one heartbeat to the beginning of the next, blood volume fluctuates throughout the body. This variation in blood volume can be observed in the skin's outer layers and measured using optical sensors [1]. Specifically, photoplethysmography (PPG) is a technology that employs LED light source and a photodetector. The LED emits light into the tissue's microvascular bed, and the photodetector, a light-sensitive sensor, records the amount of light absorbed or reflected. The light absorbed or reflected changes based on blood volume [1] [2] as seen in Figure 2.2. Heart Rate Variability (HRV) can be estimated from the photoplethysmography signal, which is equivalent to the distance between consecutive R-peaks of the ECG signal. Section 4.2.1 offers a detailed explanation of the features that can be extracted from HRV.

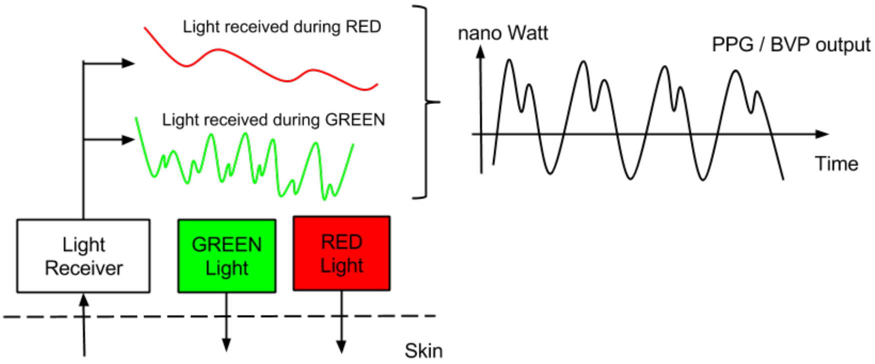


Figure 2.2: Visual Representation of PPG Signal

2.1.3 Electrodermal Activity (EDA)

Electrodermal activity (EDA), also referred to as Galvanic Skin Response (GSR) or Skin Conductance (SC), is the alteration of the skin's electrical properties due to sweating. It reflects a person's emotional state or arousal. Skin conductance variations can be measured non-invasively by applying an electrical potential between two points on the skin and measuring the current flow between them. EDA is associated with emotional arousal, and its clinical applications span a wide range of topics, including pain assessment [13] [3].

Section 4.2.2 provides a more detailed explanation of how the EDA signal is formed and the features that can be extracted from it.

2.1.4 Inter-Beat Interval (IBI)

The inter-beat interval (abbreviated as IBI) is the time interval between individual beats of the heart. It is used to estimate the instantaneous heart rate.

2.1.5 Blood Volume Pulse (BVP)

The BVP is the Blood Volume Pulse, and it's the primary output from the PPG sensor.The signal is obtained from the PPG sensor by a proprietary algorithm which combines the light signals observed during both green and red exposure as seen in Figure 2.2. It has a fixed sampling rate of 64 Hz (64 times per second).

2.1.6 Acceleration (ACC)

The 3-axis accelerometer is a sensitive sensor that measures acceleration in three dimensions (X, Y, and Z axes). It allows the device to capture a user's motion-related data, such as movements, orientation, and activity levels. This information is then used to analyze and interpret various activities, including walking, running, and other physical activities.

2.1.7 Temperature (TEMP)

The Infrared Thermopile censor reads peripheral skin temperature.The average peripheral skin temperature in humans is significantly lower than core body temperature and is subject to greater variability. It generally ranges between 25 to 35 degrees Celsius, depending on factors such as ambient temperature, local blood flow, and the individual's level of physical activity [14].

2.2 Machine Learning Algorithms

Machine Learning algorithms can be categorized into three distinct types: supervised, unsupervised, and reinforcement learning [15]. In supervised learning, each data point comes with an associated label that signifies the intended outcome, allowing the algorithm to identify patterns and generalize to situations not found in the dataset. This approach is often referred to as “learning with a teacher”. On the other hand, unsupervised techniques do not have predefined outputs for each input, so these algorithms focus on uncovering shared characteristics within the input data. Lastly, reinforcement learning operates like “training under the guidance of a judge, as the algorithm receives rewards for correct results and penalties for incorrect ones.

In this thesis, the investigated algorithms fall under the supervised learning umbrella, as every dataset sample has been labeled by a human, specifically a psychologist. The focus will be on decision tree-based algorithms, which have demonstrated strong performance in comparable tasks, as seen in previous studies [12] [5]. Decision trees are efficient classifiers resembling tree-like structures, where internal nodes represent decision tests on input variables and outgoing edges signify test outcomes. Class labels are contained within leaf nodes [16].

This subsection delves into the machine learning algorithms employed for classification purposes.

2.2.1 AdaBoost Algorithm

The AdaBoost Algorithm [17], also known as Adaptive Boosting, is an ensemble learning method that combines multiple weak classifiers to create a strong classifier.

It starts with a given dataset, assigns weights of $\frac{1}{N}$ to each instance, where N is the total number of samples, and trains a weak classifier on the weighted training data. The algorithm calculates the error rate of the weak classifier, computes its weight based on the error rate, and updates the instance weights accordingly. The weights of misclassified instances are increased, making the next classifier focus more on them. After normalizing instance weights, the process is repeated for a predetermined number of iterations or until a stopping criterion is met. Finally, the weak classifiers are combined using a weighted majority vote to form a strong classifier, which is then evaluated on a test set to measure its performance.

2.2.2 Gradient Boosting Decision Tree

Gradient Boosting Decision Tree (GBDT) [18] represents a collective learning approach that constructs numerous shallow decision trees in a sequential manner to develop a strong model. By employing gradient descent, the technique refines the model as it trains the trees based on residuals from preceding trees, which helps rectify errors and enhance overall efficacy. The end result is a model that combines the weighted sum of all the trees, offering superior accuracy compared to any single tree.

2.2.3 Bagging Decision Tree

Bagging Decision Tree [19] is an ensemble learning technique that combines multiple decision trees, trained on bootstrapped subsets of the dataset, to create a more accurate and stable model. By aggregating their predictions, the ensemble reduces overall variance and mitigates overfitting.

2.2.4 Random Forest

Random Forest [20] is an ensemble learning algorithm that constructs multiple decision trees and combines their predictions to achieve higher accuracy and stability. It improves upon Bagging Decision Trees by introducing additional randomness during the tree-building process. For each tree, a bootstrapped subset of the dataset is used, and at each node, a random subset of features is considered for splitting. This combination of random data samples and random features creates diverse trees, reducing both variance and the risk of overfitting, resulting in a more robust model.

2.2.5 Extra Trees

Extra Trees [21] is a collective learning technique that builds numerous decision trees to generate a more precise and stable model. This method amplifies the randomness already present in the Random Forest algorithm. In Extra Trees, each tree utilizes a bootstrapped subset of the dataset and a random subset of features at every node. Additionally, the algorithm chooses random split points for the features instead of the ideal ones. This heightened randomness minimizes the likelihood of overfitting, resulting in an ensemble model that boasts reduced variance and increased diversity among the individual trees.

2.3 Model Evaluation

In order to identify the best classification algorithm for the subject matter of this thesis, it is crucial to choose the most appropriate evaluation methodology and performance metrics to compare the potential algorithms.

2.3.1 Evaluation Methodology

The evaluation methodology used to compare the performance of the Machine Learning algorithms is Stratified k-fold cross-validation [22], which was used in related works in the past [4] [7] [5] [6]. Stratified k-fold cross-validation first distributes the data so that each class is proportionally represented in every fold. The dataset is then divided into k equal-sized folds, maintaining the same class label proportions as in the original dataset. Throughout k iterations of training and validation, one fold serves as the validation set while the remaining (k-1) folds form the training set. The model is trained on the training set and assessed on the validation set, producing performance metrics (such as accuracy, precision, and recall) for each iteration. Ultimately, the average of these metrics over all k iterations is computed to offer a comprehensive evaluation of the model's performance.

2.3.2 Performance Metrics

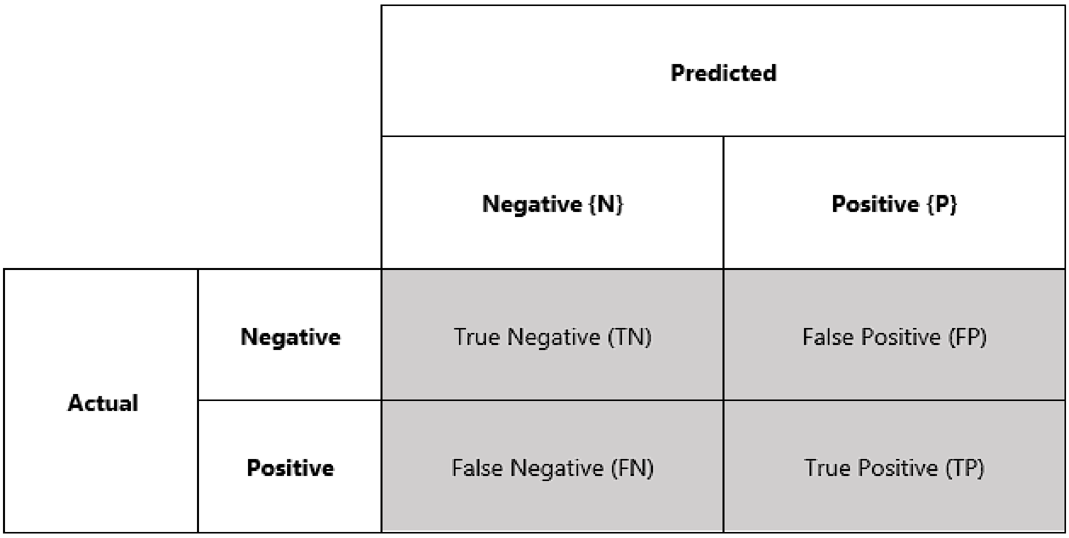
To compute the performance metrics [23], four important measures are needed. These are true positives, false positives, true negatives and false negatives, where:

- True Positives (TP): The number of samples correctly classified as positive.
- False Positives (FP): The number of samples incorrectly classified as positive.
- True Negatives (TN): The number of samples correctly classified as negative.
- False Negatives (FN): The number of samples incorrectly classified as negative

In this thesis, by positive it is meant that the participant is in the category of dysfunctional, while negative means that the participant is considered as functional. Additionally, FN are more vital – in the context of this thesis – than FP. The data are related to health care and diagnosis. Thus, it is much more important to not classify an individual as functional when in reality they are dysfunctional, because this could result to delay in diagnosis and in receiving the necessary treatment. In the opposite case, if an individual is incorrectly classified as dysfunctional, they would undergo further examinations before starting the treatment and medication, where the would possibly be correctly diagnosed.

Confusion Matrix

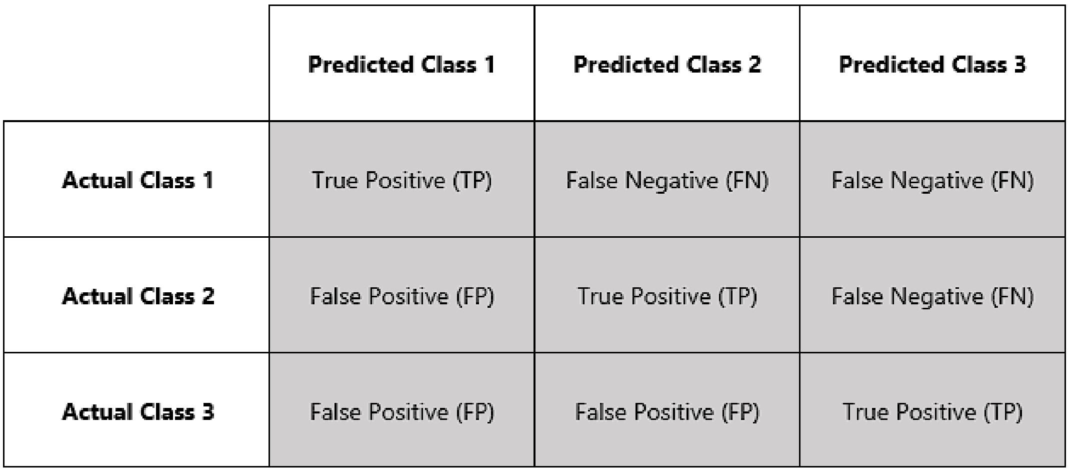
The confusion matrix is a square matrix that includes the four measures explained above. The format of a confusion matrix is shown in Figure 2.3



		Predicted	
		Negative {N}	Positive {P}
Actual	Negative	True Negative (TN)	False Positive (FP)
	Positive	False Negative (FN)	True Positive (TP)

Figure 2.3: Binary Confusion Matrix format

A variaton of the confusion matrix (Figure 2.4) is also provided as some datasets that will be studied contain three classes



	Predicted Class 1	Predicted Class 2	Predicted Class 3
Actual Class 1	True Positive (TP)	False Negative (FN)	False Negative (FN)
Actual Class 2	False Positive (FP)	True Positive (TP)	False Negative (FN)
Actual Class 3	False Positive (FP)	False Positive (FP)	True Positive (TP)

Figure 2.4: 3-Class Confusion Matrix format

Accuracy

Correct predictions to total predictions ratio.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}.$$

PPV (Precision)

True positives to total predicted positives ratio.

$$\text{PPV} = \frac{\text{TP}}{\text{TP} + \text{FP}}.$$

NPV

True negatives to total predicted negatives ratio.

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}}.$$

Specificity

True negatives to total negatives in the data ratio.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}.$$

Sensitivity (Recall)

True positives to total positives in the data ratio.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

F1-score

The harmonic mean of precision and recall

$$\text{F1-score} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

AUC

Area Under the Curve usually refers to the area under the precision-recall curve (Figure 2.5). A high AUC value implies a high-quality classification.

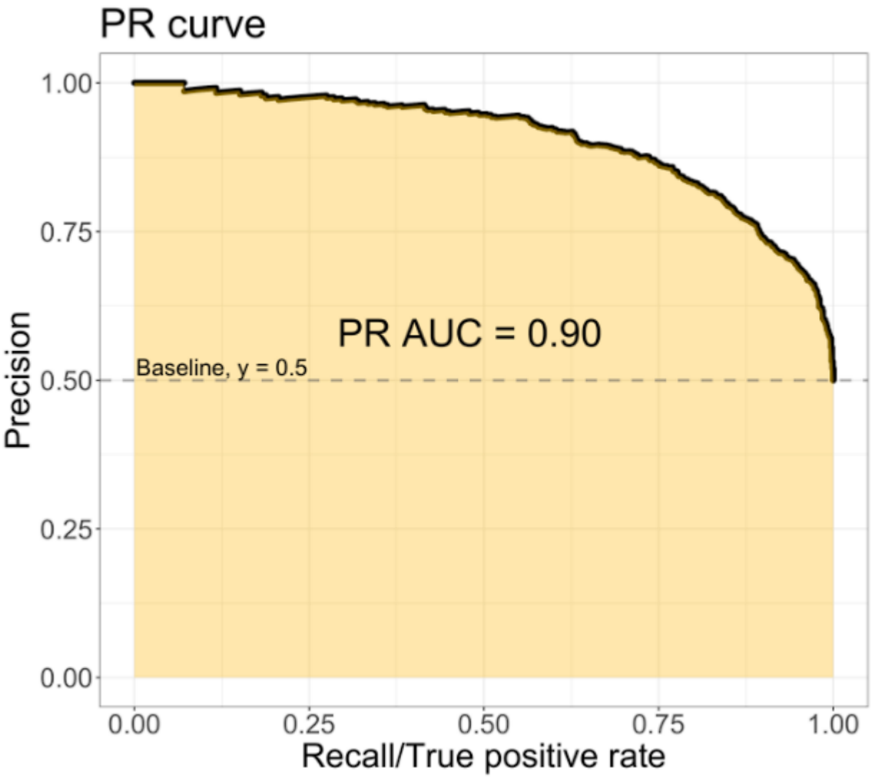


Figure 2.5: Precision-Recall Curve and AUC example

2.4 Monitoring Device - Empatica E4

In the experiment presented in Section 3, the monitoring device that was used was the Empatica E4 wristband. The Empatica E4 wristband [24] is a wearable device designed to collect and analyze real-time physiological data, making it an invaluable tool for researchers and healthcare professionals (seen in Figure 2.6). It is lightweight as it only weights 25g, has a streaming mode of 24+ hours as well as memory mode of 48+ hours [25]. Equipped with a photoplethysmography (PPG) sensor, the E4 measures blood volume pulse to derive heart rate, heart rate variability, and other cardiovascular parameters. Additionally, the wristband incorporates an electrodermal activity (EDA) sensor, which measures the electrical conductance of the skin to provide insights into emotional arousal, stress, and various psychological states. The device also features a 3-axis accelerometer to capture information on physical activity, movement, and gestures, as well as an infrared thermopile to measure skin temperature for studying thermal regulation and other physiological parameters [26] [27].

The Empatica E4 wristband has been widely utilized in diverse research settings, including stress monitoring, sleep studies, emotion recognition, and mental health research, making it a versatile and powerful tool for both clinical trials and academic research projects. The Empatica E4 wristband was chosen for the ongoing study because of its wide range of sensors and demonstrated effectiveness in various research environments. This ensures precise, dependable data gathering and the capability to examine multiple physiological parameters simultaneously.

Figure 2.6: Real Look at the Empatica E4 wristband

Chapter 3

Data Collection and Previous Work

Contents

3.1 Psysiological Experiments

The University of Cyprus' Department of Psychology conducted a series of four experiments: Diagnosis of Experiential Avoidance in Smokers, Diagnosis of Eating Disorders, Diagnosis of Experimental Avoidance for Anxiety, and Functional versus Dysfunctional Coping with Acute Pain. These experiments took place in the ACTHealth lab, and participants were volunteers. This section outlines the procedure for each experiment, as well as the methodologies of related works [4] [12] [5] .

All experiments are connected to Acceptance and Commitment Therapy, where individuals are divided into two groups (acceptance or avoidance) based on their reactions. As described in Section 1.2, acceptance-based strategies involve individuals accepting their thoughts and sensations (functional), while avoidance-based strategies are associated with avoiding uncomfortable thoughts and sensations or attempting to control, alter, or avoid them (dysfunctional) [33]. Additionally, a person's classification can change depending on the environment and circumstances. During the data collection for the first three experiments, it was assumed that each participant belonged to a single group throughout the entire procedure. However, this hypothesis did not apply to the fourth experiment.

3.1.1 Diagnosis of Experiential Avoidance in Smokers

The goal of this experiment was to compare the emotional regulation abilities between smokers in the acceptance category and those in the avoidance category. The experiment consisted of five consecutive timeframes, each lasting 8 minutes. The first timeframe was used as a baseline and to ensure that the participant was in a calm state. In the next two timeframes, an emotionally neutral video was shown, and during the final two timeframes, the participant viewed a video designed to evoke negative emotions. Using the collected data, an expert from the Department of Psychology classified the participant into one of the categories.

During the entire procedure, the signals recorded from the participants were ECG, COR (using fEMG), and GSR with a sampling rate of 1000Hz. Throughout the latter four timeframes, participants were asked to complete a series of cognitive tests, which are short assessments of how effectively the brain functions.

3.1.2 Diagnosis of Eating Disorders

The aim of this experiment was to compare the emotional regulation abilities between people with low risk and high risk of having an Eating Disorder. The experiment consisted of five consecutive timeframes, each lasting 2.5 minutes. The first timeframe was used as a baseline and to ensure that the participant was in a calm state. In the second and fourth timeframes, an emotionally neutral video was shown. In the third timeframe, the participant viewed an unpleasant general-content video, while in the fifth timeframe, the participant viewed an unpleasant video related to eating disorders.

Using the collected data, an expert from the Department of Psychology classified the participants into one of the categories. The signals recorded from the participants during the entire procedure were ECG, COR (using fEMG), and GSR with a sampling rate of 1000Hz. The participants were also asked to complete the Body Image Acceptance and Action Questionnaire (BI-AAQ), which measures body image flexibility. Participants responded on a seven-point scale from never true to always true, where higher summed scores indicate greater body image flexibility [28].

3.1.3 Diagnosis of Experiential Avoidance for Anxiety

The aim of this experiment was to compare the emotional regulation abilities between people who belong to the acceptance category and people who belong to the avoidance category regarding anxiety. The experiment consisted of 72 consecutive timeframes, each lasting around 1.8 minutes. In each timeframe, the participant was shown a single image, which was supposed to cause a different reaction depending on whether they showed signs of anxiety or not. Using the collected data, an expert from the Department of Psychology classified the participant into one of the categories.

The signals that were recorded from the participants during the entire procedure were ECG, GSR, and fEMG (COR, ORB, and ZYG muscles) with a sampling rate of 1000Hz.

3.1.4 Functional Versus Dysfunctional Coping with Acute Pain

In this study, 80 people participated [29], with the aim of comparing acceptance and avoidance coping strategies in a pain-induction experiment. Participants were randomly split into four groups (conditions), with each group receiving different instructions on how to deal with pain. The four conditions were: (a) Acceptance followed by avoidance; (b) Avoidance followed by acceptance; (c) No instructions given (control) followed by acceptance; and (d) No instructions given (control) followed by avoidance.

The experiment consisted of three timeframes. The first timeframe lasted 5 minutes and served as a baseline to ensure that the participant was in a state of calm. Next, participants were instructed on how to deal with pain in the following timeframe, based on their condition. The second timeframe followed, during which participants were subjected to the Cold Pressor Task (CPT) – immersing their hand in a container filled with cold water for as long as they could. Afterward, participants were instructed on how to deal with pain in the last timeframe, based on their condition. The third timeframe followed, during which participants were subjected to a second CPT. The maximum duration of the second and third timeframes was 3 minutes.

Multiple measures – behavioral, psychophysiological, and self-reported – were recorded throughout the entire procedure. Behavioral measures include pain threshold and pain tolerance, which are the number of seconds that elapsed from immersion until the participant verbally reported pain and until the participant removed their hand from the container, respectively. The psychophysiological signals collected using the stationary device were ECG, part of the EDA signal (SCL, as explained in Section 4.2.2), and fEMG (COR and ZYG muscles), with sampling frequencies of 1kHz, 250Hz, and 1kHz, respectively. The measures collected using the band were PPG and EDA, with sampling frequencies of 1Hz and 0.2Hz, respectively. The only measure collected from the ring was EDA, with a sampling frequency of 3Hz. As for self-reported data, participants completed several questionnaires examining various aspects, including their psychological condition and their use of pain-coping strategies.

3.1.5 Functional Versus Dysfunctional Coping in Real Time

This is the most recent experiment, and the one that this thesis is focuses on. Participants entered the Ecological Momentary Assessment (EMA) phase [30]. In this phase, they were given smartphones and wearable psychophysiological monitors to wear for the next three days.Participants were instructed on how to wear the Empatica E4 wristband, as well as how to charge the device and ensure data collection. They were prompted to respond to questions on an app pre-installed by the researchers on the provided smartphones. Specifically, participants answered questions five times a day, at fixed intervals throughout the three days, every three hours from 10am to 10pm. Participants wore the monitors until bedtime, at which point they charged them. The participants were asked about social context, experiences of stress or pain (both physical and emotional), and the use of coping strategies. After the three days, participants returned the devices. If participants did not respond, reminder messages were automatically sent every 30 minutes.

3.2 Related Work

Four prior works examined the data of the experiments regarding smoking, eating disorders, anxiety as well as pain and emotions management. The methodologies of these works are presented in the current section.

This thesis concentrates on an experiment that investigates pain and emotion management in real time mentioned in Section 3.1.5, with the work carried out being detailed in Chapters 4 through 6.

3.2.1 Diploma Project of Ch. Galazis in 2017

The initial analysis [4] aimed to identify the best feature combination for classifying experiments related to smoking and eating disorders, utilizing knowledge from prior research [7]. Additional work was conducted for the anxiety experiment. All unique feature combinations were used to train and test the Random Forest classifier, with candidate features being the mean values of each recorded signal in each timeframe.

The chosen combination had the highest accuracy and the fewest features. The machine learning algorithms studied included Logistic Regression, Naive Bayes, K-Nearest Neighbours, Classification Tree, Neural Network, SVM, Bagging (using Decision Tree as the Base Learner), AdaBoosting (using Decision Tree as the Base Learner), Gradient Tree Boosting, and Random Forest. The data were divided into training and test sets in ten different ways, and each algorithm was executed ten times, using the results from the best-performing distribution for algorithm comparison.Results of the work as well as subsequent works will be presented in Section 6.3

3.2.2 Master Thesis of A. Trigeorgi in 2018

A more recent study [12] employed a different approach, with more emphasis on feature extraction. Time-domain features were extracted from the ECG signal, resulting in candidate features that included not only the mean values of each signal but also

ECG-derived time-domain features (explained in detail in Chapter 4.1.1). To select the optimal feature combination, a Random Forest Classifier was used with Stratified k-fold cross-validation, averaging the performance across k iterations.

The same algorithms from the previous study were examined, but the execution method differed, using Stratified 5-fold cross-validation. The data were divided into training and test sets in five different ways, and each algorithm was executed five times, with the average performance of the five runs measured.

3.2.3 Master Thesis of G. Demosthenous in 2019

The later study [5] extracted even more features from the ECG signal. To identify the most effective feature combination, Breiman and Friedman's method [31] was employed to calculate feature importance using Gradient Boosting Decision Tree, ranking candidate features based on node impurity.

The algorithms studied diverged from the previous two studies, focusing on tree-based algorithms. The five algorithms analyzed were Gradient Boosting Decision Tree (GBDT), Ada Boosting Decision Tree, Bagging Decision Tree (BDT), Random Forest (RF), and Extra Trees (ET). An additional step, training data multiplication, was performed to increase the sample size and counter the assumption that each participant belonged to the same group throughout the experiment. Two methodologies were used and compared: Moving Window Methodology (MWM) and Rectangular Window Methodology (RWM) [32]. The algorithm execution method combined the methods used in the previous two studies, using 10-fold cross-validation and executing each algorithm 10 times for each split, totaling 100 executions per algorithm.In this work, prediction was also made for samples from previous experiments [33].

3.2.4 Diploma Project of E. Georgiou in 2022

This thesis [6] expanded on the previous works. The study involved 80 participants undergoing the Cold Pressor Task, a pain threshold test. The participants were monitored using three devices: BIOPAC, Microsoft Band 2, and the Moodmetric Smart Ring, which recorded various physiological signals at different frequencies. Given the relatively small dataset, artificial samples were created using the Rectangular Window Methodology, generating four datasets of different window sizes (10, 20, 30, and 40 seconds). Various features were extracted from each signal, with time-domain measures being the main focus for ECG and HRV signals. The EDA signal was divided into SCL and SCR components, from which statistical metrics were extracted. The primary focus of the thesis was on the selection of the most relevant features. This was done using three feature selection methods: Wrapper, Embedded, and Filter Methods. The thesis also compared the common signals from different monitoring devices. The findings varied depending on the methods used for comparison. Lastly, the thesis found that data multiplication using Rectangular Window Methodology improved classifier performance and data from the Microsoft Band 2 could match the performance of stationary devices.

Chapter 4

Signal Analysis

Contents

4.1 Data Selection

A comprehensive data collection process for machine learning was conducted. The study (mentioned in Section 3.1.5) involved a total of 88 participants, with each participant having five files associated with them. These files contained various physiological metrics, namely Acceleration (ACC), Interbeat Interval (IBI), Heart Rate Variability (HRV), Temperature (TEMP), and (EDA). The data spanned a period of three days for each patient.

In addition to the metrics, a questionnaire was completed by each participant five times a day, with the time of completion being recorded. The primary question in the questionnaire was “What are you doing right now to manage your thoughts, feelings and emotions?”. Three different answers were provided for patients to choose from, labeled as **avoidance** for ”I distract myself by doing or thinking about something else so that I avoid thinking about them ”, **acceptance** for “I let the unpleasant thoughts and experiences be there without doing anything to drive them away”, and **mindfullness** for “I focus on what I’m doing now”.

The first step was to go through each questionnaire and identify the time at which the pain-coping question was answered by the participant. Then, the respective metric files were accessed, and the corresponding time was navigated to. The key data points extracted for analysis were the values recorded five minutes before and five minutes after the question was answered by the participant. This approach allowed for a focused examination of the relationship between physiological metrics and patients’ pain-coping strategies.

4.2 Feature Extraction

Psychophysiological signals in their raw form are unsuitable for efficiently training algorithms. As a result, it is necessary to derive significant features from these signals. This section provides a detailed explanation of the extracted features for each signal.

4.2.1 Features Extracted from PPG Signal

Time-domain measures (also called HRV time-domain measures) are based on Heart Rate Variability, which is known as RR intervals. An RR interval is the time elapsed between two consecutive R peaks. RR intervals are also referred to as NN intervals. In fact, NN intervals are the time elapsed between two consecutive normal R peaks, which are the R peaks that do not include artifacts [8]. To extract the features from the raw signals derived from the PPG Signal, the flirt [34] Python library was used.Firstly RR intervals are computed, as shown in Table 4.1, as well as the differences (RRdiff) and squared differences (RRsqdiff) between consecutive RR intervals.

Abbreviation	Explanation	Formula
RR	As explained above, it is the interval of consecutive R waves in milliseconds.	$RR = \frac{diff(R_{peaks})}{sf} \times 1000$, where sf is the sampling frequency
RRdiff	The absolute value of the differences between consecutive RR intervals	RR$_{diff} = \|diff(RR)\|$
RRsqdiff	The squared differences of consecutive RR intervals	RR$_{sqdiff} = RR_{diff}^2$

Table 4.1: Measures used to express time-domain-features

Based on the above, the time-domain features of Table 4.2 can be derived.

Abbreviation	Explanation	Formula	Unit
IBI	Inter-Beat Intervals. The average of RR intervals, it is the interval of consecutive R waves in milliseconds.	$IBI = \overline{RR}$	ms
BPM	Beats Per Minute. The average number of heart beats per minute.	$BPM = \frac{60000}{\overline{RR}}$	bpm
SDNN	Standard Deviation of NN intervals	$SDNN = \sqrt{\frac{1}{N-1}\sum(RR_i - \overline{RR})^2}$	ms
SDSD	Standard Deviation of Successive Differences between consecutive RR intervals	$SDSD = \sqrt{\frac{1}{N-1}\sum(RR_{diff_i} - \overline{RR_{diff}})^2}$	ms

Continued on next page

Table 4.1 Continued from previous page

Abbreviation	Explanation	Formula	Unit
RMSSD	Root Mean Square of Successive RR interval Differences	$RMSSD = \sqrt{\frac{1}{N-1}\sum(RR_{diff_i})^2}$	ms
pNN20	The ratio of differences of consecutive NN intervals that are greater than 20ms to all consecutive NN intervals	$pNN20 = \frac{count(diff(RR)>20ms)}{count(diff(RR))}$, where count(X) gives the number of elements in X	%
pNN50	The ratio of differences of consecutive NN intervals that are greater than 50ms to all consecutive NN intervals	$pNN50 = \frac{count(diff(RR)>50ms)}{count(diff(RR))}$, where count(X) gives the number of elements in X	%
HRMAD	The Median Absolute Deviation of the Heart Rate	$HR_{mad} = median(\|RR_i - \tilde{RR}\|)$, where $\tilde{RR} = median(RR)$, where count(X) gives the number of elements in X	bpm

Table 4.2: Measures used to express time-domain-features

In addition some frequency domain features were extracted as shown in Table 4.3.These features are derived from the power spectral density (PSD) of the RR intervals. The PSD is calculated using methods like Fast Fourier Transform (FFT) or autoregressive methods. The total power of the PSD is divided into different frequency bands.

Abbreviation	Explanation
VLF (Very Low Frequency)	Power in the very low frequency band (typically 0.0033 to 0.04 Hz).
LF (Low Frequency)	Power in the low frequency band (typically 0.04 to 0.15 Hz).
HF (High Frequency)	Power in the high frequency band (typically 0.15 to 0.4 Hz).
LF/HF Ratio	The ratio of LF power to HF power.
LFnU (LF normalized units)	LF power in normalized units, which is LF power divided by the total power minus VLF power, and then multiplied by 100.
HFnU (HF normalized units)	HF power in normalized units, which is HF power divided by the total power minus VLF power, and then multiplied by 100.

Table 4.3: Measures used to express frequency-domain-features

Moreover, non-linear features are derived from the Poincaré plot, which is a scatterplot of the current RR interval against the next RR interval, as shown in Table 4.4

Feature Name	Feature Description
SD1	The standard deviation of points perpendicular to the line of identity on the Poincaré plot. It measures the short-term variability of heart rate.
SD2	The standard deviation of points along the line of identity on the Poincaré plot. It measures the long-term variability of heart rate.
SD2/SD1 Ratio	The ratio of SD2 to SD1.

Table 4.4: Measures used to express non-linear features

4.2.2 Features Extracted from EDA Signal

There are four types of features that can be extracted from EDA. These are time domain features, frequency domain features, time-frequency domain features, and Mel-frequency cepstrum features [35].This thesis focuses on statistical features, as they are widely used in the literature.The EDA signal can be decomposed into two main components, the Tonic Component (Skin Conductance Level, SCL) which is the slow-changing, baseline level of skin conductance. It represents the overall level of arousal or stress over a longer period of time.The second is the Phasic Component (Skin Conductance Response, SCR) which is the fast, transient changes in skin conductance in response to specific events or stimuli. It represents the immediate or short-term response to a stimulus or event.To extract the features shown in Table 4.5, the flirt [34] library was utilised.

Feature Name	Feature Description
tonic_mean	Average value of SCL.
phasic_mean	Average value of SCR.

Table 4.5: Statistical features extracted from EDA

4.2.3 Features Extracted from 3-axis Accelerometer

Lastly, the features extracted from the 3-axis are shown in Table 4.6.The signal [27] generally measures proper acceleration relative to freefall.In the context of human activity monitoring, it is used to measure movements of the body.Once again, the flirt [34] Python library was used to extract those features.

Feature Name	Feature Description
acc_x_mean	Average acceleration along X-axis.
acc_y_mean	Average acceleration along Y-axis.
acc_z_mean	Average acceleration along Z-axis.

Table 4.6: Features extracted from ACC

Chapter 5

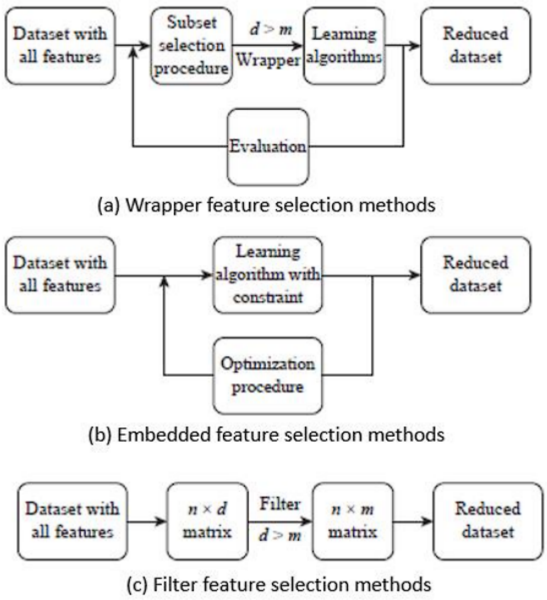
Feature Selection

Contents

5.1 Feature Selection Techniques

Feature selection is a crucial part of the data preprocessing pipeline, as it greatly enhances the performance and comprehensibility of machine learning models. By identifying the most relevant features, the dimensionality of the dataset can be reduced, which decreases computational complexity and processing time. This reduction allows models to train more quickly and possibly achieve better generalization by reducing the likelihood of overfitting to irrelevant or noisy features. By recognizing the influence of each feature on the model's predictive capability, more informed decisions can be made, potentially uncovering new insights that lead to more effective and accurate outcomes.

In the context of supervised data, a variety of feature selection techniques are available. Specifically, there are three main categories: Wrapper, Filter and Embedded Methods. Those three main categories will be examined in order to compare their results. Figure 5.1 shows the main concept of the categories explained in more detail in this chapter.



(a) Wrapper feature selection methods

(b) Embedded feature selection methods

(c) Filter feature selection methods

Figure 5.1: Flowchart of the three main feature selection methods

Before using the aforementioned methods, the Variance Inflation Factor (VIF) technique was utilized [36] to address the issue of multicollinearity and to drop unwanted features from the dataset . Multicollinearity occurs when two or more predictor variables are highly correlated, which can lead to unstable estimates and reduced interpretability of the model. The VIF method quantifies the severity of multicollinearity by measuring the extent to which the variance of a regression coefficient is inflated due to the presence of correlated features. A high VIF value for a given feature indicates that the feature is highly correlated with other features in the dataset, and thus, it may be redundant or provide limited additional information. By calculating the VIF for each feature and removing those with VIF values above a predetermined threshold,multicollinearity was effectively reduced and improved the overall performance and interpretability of the predictive model. This approach helped retain only the most relevant and informative features, contributing to a more reliable and robust model.

5.1.1 Wrapper Methods

Wrapper methods for feature selection are a set of techniques used in machine learning to identify the most relevant features for a given predictive model. These methods focus on searching for the optimal subset of features by evaluating their contribution to the model's performance [37]. The key technique involves building and assessing multiple models with different feature combinations, and selecting the one that yields the best performance based on a specific evaluation metric. Wrapper methods can employ approaches such as forward selection, backward elimination, or recursive feature elimination to generate feature subsets.

In this study, the RandomForestClassifier from the scikit-learn library in Python was employed to identify the best set of features for the predictive model. The dataset was first loaded and preprocessed, with the target variable separated from the predictor variables.A RandomForestClassifier was trained and the built-in `feature_importances_` attribute was utilized to compute the importance of each feature. These feature importances were ranked in descending order, and the selected features were utilized subsequently.By leveraging the feature importance attribute of RandomForestClassifier, we effectively identified the most informative set of features, which contributed to improving the overall accuracy of our predictive model.

5.1.2 Filter Methods

In recent years, filter methods have gained significant attention for feature selection in Machine Learning, primarily due to their computational efficiency and simplicity. These methods rely on evaluating the intrinsic properties of the dataset, such as correlation and mutual information, to determine the relevance and usefulness of features. By eliminating redundant or irrelevant features, filter methods reduce the dimensionality of the dataset, thereby improving the performance of machine learning models [37]. The adoption of filter methods not only helps to mitigate overfitting but also enhances the interpretability and generalization capabilities of the models, which are crucial for robust and reliable applications.

In this thesis, the SelectKBest method was employed, a filter-based feature selection technique, to identify the most relevant features from the dataset. SelectKBest works by ranking features according to their scores, which are calculated using a univariate statistical test, such as chi-square, ANOVA F-value, or mutual information. By selecting the top K features with the highest scores, only the most informative and relevant featureswere retained. This approach led to a significant reduction in computational complexity but also resulted in improved model performance and generalization, demonstrating the effectiveness of SelectKBest in the context of this research.

The `f_classif` scoring function within the SelectKBest method was used, as it is designed to handle continuous features with a categorical target. The `f_classif` function is based on the one-way ANOVA F-value, which computes the variance between the group means and within-group variances, thereby quantifying the extent to which a particular feature can discriminate between different target classes. By selecting the top K features with the highest F-values, the most significant and relevant features were retained, leading to a more efficient and generalizable prediction performance.

5.2 Selected Features for the Initial Dataset

In this section, the selected features from the feature selection process will be presented. This process was applied to four different variations of the initial dataset. These variations include a) the original 3-class dataset, b) the dataset with ”mindfulness” treated as ”avoidance”, resulting in 2 classes - ”acceptance” and ”avoidance” ,c) the dataset with ”mindfulness” treated as ”acceptance,” also resulting in 2 classes - ”acceptance” and ”avoidance” and d) the dataset with all samples of ”mindfulness” removed, leaving 2 classes - ”acceptance” and ”avoidance.” The feature selection process employed RandomForestClassifier and SelectKBest techniques for each of these datasets. Following this introduction, a table that summarizes the selected features for each dataset variation will be provided, offering insights into the differences in feature importance across the different dataset structures.

	RandomForestClassifier	SelectKBest
Original 3-class dataset	'hrv_mean_nni', 'hrv_sdnn', 'hrv_pnni_50', 'hrv_pnni_20', 'hrv_lf', 'hrv_hf', 'phasic_mean', 'acc_x_mean', 'acc_y_mean', 'temp_mean', 'ibi', 'sdnn', 'pnn20', 'sd2', 'sd1/sd2', 'hf_perc', 'hf_nu'	'hrv_rmssd', 'hrv_sdsd', 'hrv_sdnn', 'hrv_pnni_50', 'hrv_pnni_20', 'hrv_lf', 'hrv_hf', 'temp_mean'
”Mindfulness” as ”avoidance”	'hrv_mean_hr', 'hrv_rmssd', 'hrv_sdnn', 'hrv_pnni_50', 'hrv_pnni_20', 'hrv_lf', 'hrv_hf', 'hrv_lf_hf_ratio', 'acc_x_mean', 'acc_z_mean', 'temp_mean', 'bpm', 'sdnn', 'sdsd', 'sd2', 's', 'sd1/sd2'	'hrv_pnni_50', 'hrv_pnni_20', 'hrv_hf', 'temp_mean', 'bpm', 'ibi', 'pnn20', 'pnn50'
”Mindfulness” as ”acceptance”	'hrv_mean_hr', 'hrv_sdsd', 'hrv_sdnn', 'hrv_pnni_50', 'hrv_pnni_20', 'hrv_lf', 'hrv_hf', 'hrv_lf_hf_ratio', 'tonic_mean', 'phasic_mean', 'acc_y_mean', 'temp_mean', 'bpm', 'ibi', 'sdnn', 'lf/hf', 'lf_nu'	'hrv_mean_hr', 'hrv_mean_nni', 'hrv_rmssd', 'hrv_sdsd', 'hrv_sdnn', 'hrv_lf', 'hrv_hf', 'lf/hf'

Continued on next page

Table 5.0 (*continued from previous page*)

	RandomForestClassifier	SelectKBest
”Mindfulness” removed	`'hrv_rmssd', 'hrv_sdsd', 'hrv_sdnn', 'hrv_pnni_50', 'hrv_lf', 'hrv_hf', 'hrv_lf_hf_ratio', 'tonic_mean', 'acc_x_mean', 'acc_y_mean', 'temp_mean', 'bpm', 'ibi', 'sdnn', 'sdsd', 'sd1/sd2', 'breathingrate'`	`'hrv_rmssd', 'hrv_sdsd', 'hrv_sdnn', 'hrv_lf', 'hrv_hf', 'temp_mean', 'bpm', 'ibi'`

Table 5.1: Selected Features for original dataset

5.3 Selected Features for the Optimized Dataset

In this section, the need to rerun the feature selection process due to changes in the dataset after cleaning and optimization will be discussed. The data was cleaned to provide better samples for the classification task, thus enhancing the reliability of the analysis. The specific criteria for removing samples and the detailed process of data cleaning will be elaborated upon in Chapter 6. The impact of these changes on the feature selection process, as well as any differences in the selected features between the initial and optimized datasets, will be examined in this section. Following the discussion, a table containing the selected features for the optimized dataset will be provided, allowing for a comparison with the features selected from the initial dataset.

	RandomForestClassifier	SelectKBest
Original 3-class dataset	'hrv_mean_hr', 'hrv_pnni_50', 'hrv_pnni_20', 'hrv_hf', 'hrv_lf_hf_ratio', 'tonic_mean', 'phasic_mean', 'acc_x_mean', 'acc_y_mean', 'acc_z_mean', 'temp_mean', 'sdnn', 'sdsd', 'hr_mad', 'sd2', 's', 'sd1/sd2'	'hrv_hf', 'acc_x_mean', 'temp_mean', 'pnn20', 'lf/hf', 'vlf_perc', 'hf_perc', 'lf_nu'
”Mindfulness” as ”avoidance”	'hrv_pnni_50', 'hrv_pnni_20', 'hrv_lf', 'hrv_hf', 'hrv_lf_hf_ratio', 'tonic_mean', 'phasic_mean', 'acc_y_mean', 'temp_mean', 'sdnn', 'sdsd', 'rmssd', 'hr_mad', 'sd1', 'sd2', 's', 'sd1/sd2'	'hrv_mean_nni', 'hrv_hf', 'acc_x_mean', 'temp_mean', 'bpm', 'ibi', 'pnn20', 'pnn50'
”Mindfulness” as ”acceptance”	'hrv_mean_hr', 'hrv_mean_nni', 'hrv_hf', 'hrv_lf_hf_ratio', 'tonic_mean', 'acc_x_mean', 'acc_y_mean', 'acc_z_mean', 'ibi', 'sdnn', 'hr_mad', 'sd2', 's', 'lf/hf', 'lf_perc', 'lf_nu', 'hf_nu'	'acc_x_mean', 'acc_y_mean', 'hr_mad', 'lf/hf', 'vlf_perc', 'hf_perc', 'lf_nu', 'hf_nu'

Table 5.2: Selected Features for the Optimized dataset

5.4 Data Cleaning

In order to create the optimized dataset, certain criteria were used to identify and remove invalid samples. Two main criteria guided the data cleaning process:

Temperature Mean: Samples with a mean temperature below 28 degrees Celsius were considered invalid. The rationale behind this criterion is that 28 degrees Celsius is approximately room temperature, indicating that the patient likely was not wearing the device when the signals were recorded at the time when the participant was answering the questionnaire.

Questionnaire Responses: Samples were also removed if the patients indicated in the questionnaire that they did not use any of the techniques taught in the lab experiment. These samples were considered invalid, as the aim was to include only “real” and correctly labeled samples in the analysis.

By removing samples based on these criteria, the dataset was refined to include more accurate and representative data points. The impact of this data cleaning process on the balance of the dataset and the subsequent feature selection process will be discussed in the following chapter.

Chapter 6

Classification

Contents

6.1 Classification Process

The classification process in this study plays a crucial role in determining the effectiveness of different machine learning algorithms and the impact of the optimized dataset on their performance. To achieve this, a structured methodology has been followed to ensure a rigorous analysis and a fair comparison of the algorithms and datasets.

The data preparation step is important, as it involves splitting the dataset into training and testing sets. This study will use results from both the initial and optimized datasets, ensuring that the machine learning models are trained on one subset of the data and evaluated on an unseen subset. This approach helps avoid overfitting and provides an unbiased evaluation of their performance.

Five machine learning algorithms have been selected for the classification task. These algorithms represent a diverse range of techniques, enabling a comprehensive comparison of their performance on the given task. The selected algorithms, which were explained in Section 2.2, include AdaBoost Algorithm, Gradient Boosting Decision Tree, Bagging Decision Tree, Random Forest, and Extra Trees. Each of these algorithms is trained on the training set using the selected features obtained through the feature selection process. The models are optimized by tuning their hyperparameters to achieve the best possible performance on the training data.

After training, the models are evaluated on the testing set. Various evaluation metrics (Section 2.3.2), such as accuracy, recall, specificity (true negative rate), and F1-score, are calculated to assess the performance of each classifier. The performance of the five classifiers is compared using these evaluation metrics, with the primary goal of comparing different sets of the dataset and, secondly, to compare the different algorithms. The best-performing classifier(s) are identified based on these metrics, and their suitability for detecting and preventing dysfunctional pain coping is discussed.

In summary, the classification process serves as a means to assess the effectiveness of the selected machine learning algorithms and the impact of the optimized dataset on their performance. By comparing the results from both the initial and optimized datasets, the study aims to provide valuable insights into the benefits of data cleaning and optimization in the context of detecting and preventing dysfunctional pain coping.

6.2 Classifier Selection

As we transition into the Classifier Selection chapter, we will be delving deeper into the comparative analysis of the seven different scenarios that have been thoroughly investigated in this study. A fundamental aspect of this analysis involves understanding the varying distribution of samples across these scenarios, which plays a critical role in the performance of the different machine learning algorithms utilized. In order to facilitate a comprehensive and clear understanding, Table 6.1 illustrated the number of samples for each scenario. This will serve as a foundation for the subsequent discussions and evaluations of the classifiers, thereby aiding in the selection of the most appropriate model based on the specific characteristics of each scenario.

Scenarios	Avoidance	Acceptance	Mindfullness
Original Dataset with three classes	124	79	316
Original Dataset with Mindfullness Treated as Avoidance	440	79	-
Original Dataset with Mindfullness Treated as Acceptance	124	395	-
Original Dataset with removed mindfullness	124	79	-
Optimised Dataset with three classes	12	16	192
Optimised Dataset with Mindfullness Treated as Avoidance	204	16	-
Optimised Dataset with Mindfullness Treated as Acceptance	12	208	-

Table 6.1: Distribution of Samples Across Scenarios

6.2.1 The Original Dataset with Three Classes

This section introduces the first scenario of the classification process, which involves using the original dataset containing three classes: acceptance, avoidance, and mindfulness. In this scenario, the dataset remains unaltered, and the primary goal is to evaluate the performance of the five selected machine learning algorithms - AdaBoost Algorithm, Gradient Boosting Decision Tree, Bagging Decision Tree, Random Forest, and Extra Trees - in classifying patients based on their pain coping strategies.

The aim is to assess the baseline performance of the classifiers in distinguishing between functional and dysfunctional pain coping strategies while maintaining the original structure of the dataset. By establishing this baseline, it will be possible to compare the performance of the classifiers in subsequent scenarios involving variations of the dataset, thereby shedding light on the impact of data modifications on the classification outcomes. The results of the classification process for this scenario, along with the performance metrics for each of the machine learning algorithms, will be presented and discussed in the following sections.

	GBDT		ABDT		BDT		RF		ET	
	Value	SD	Value	SD	Value	SD	Value	SD	Value	SD
Accuracy	0.64	0.05	0.57	0.05	0.66	0.05	0.67	0.04	0.68	0.04
F1-score	0.48	0.08	0.42	0.07	0.49	0.08	0.48	0.08	0.49	0.08
Recall	0.86	0.07	0.79	0.07	0.88	0.06	0.92	0.05	0.94	0.04
Precision(PPV)	0.68	0.04	0.65	0.04	0.69	0.04	0.68	0.04	0.68	0.04
AUC	0.69	0.06	0.60	0.05	0.73	0.06	0.74	0.06	0.73	0.06
Specificity	0.37	0.11	0.34	0.10	0.37	0.12	0.33	0.11	0.31	0.11
NPV	0.78	0.05	0.72	0.04	0.79	0.05	0.81	0.05	0.82	0.05

Table 6.2: 3-Class RFE

In this analysis, machine learning algorithms were evaluated on the original 3-class dataset using various performance metrics. The best-performing algorithms (Extra Trees and Random Forest), as seen from Table 6.2, generally provided superior results across most metrics, such as Accuracy (0.68 ± 0.04), F1-score (0.49 ± 0.08), Recall (0.94 ± 0.04), AUC (0.73 ± 0.06), and NPV (0.82 ± 0.05). Precision (PPV) was found to be similar across all algorithms, ranging from 0.65 to 0.69. However, specificity remained relatively low for all algorithms, potentially indicating difficulties in distinguishing between certain classes. While some algorithms consistently demonstrated lower performance in most metrics, the findings suggest that the top-performing algorithms are the most promising approaches for this dataset. Further investigation might be needed to improve specificity and overall performance.

	GBDT		ABDT		BDT		RF		ET	
	Value	SD	Value	SD	Value	SD	Value	SD	Value	SD
Accuracy	0.61	0.05	0.52	0.06	0.63	0.06	0.64	0.05	0.65	0.05
F1-score	0.44	0.08	0.37	0.07	0.47	0.08	0.47	0.07	0.47	0.09
Recall	0.44	0.06	0.37	0.06	0.46	0.07	0.46	0.06	0.47	0.07
Precision(PPV)	nan	nan	0.39	0.10	nan	nan	nan	nan	nan	nan
AUC	0.67	0.07	0.58	0.06	0.69	0.07	0.70	0.06	0.70	0.06
Specificity	0.72	0.03	0.68	0.03	0.73	0.04	0.73	0.03	0.74	0.04
NPV	0.75	0.05	0.68	0.05	0.76	0.06	0.77	0.05	0.78	0.06

Table 6.3: 3-Class SelectKBest

In this analysis, the same 3-class dataset was used applying SelectKBest feature selection. The results from Table 6.3 show that the top-performing algorithms generally demonstrate improved performance in some metrics, such as Specificity (0.74 ± 0.04), NPV (0.78 ± 0.06), and AUC (0.70 ± 0.06). However, there is a decline in other metrics like Accuracy (0.65 ± 0.05), F1-score (0.47 ± 0.09), and Recall (0.47 ± 0.07) compared to the previous analysis using RFE.

It is worth noting that the Precision (PPV) metric for some algorithms is not available due to zero True Positives and False Positives, which makes the calculation of PPV impossible in these cases. The specificity has improved considerably for all algorithms compared to the previous results, indicating better performance in distinguishing between classes after feature selection.

The top-performing algorithms still appear promising for this dataset. However, the overall performance seems to have been affected by the feature selection process, particularly in terms of Accuracy, F1-score, and Recall. The lack of True Positives and False Positives for some algorithms in calculating PPV suggests that these models may struggle to correctly identify positive cases. Further investigation is recommended to understand the impact of feature selection on the dataset and to identify the best combination of algorithms and preprocessing techniques for optimal performance.

6.2.2 The Original Dataset with Mindfullness Treated as Avoidance

In this section, the second scenario of the classification process is introduced, which involves using a modified version of the original dataset. In this scenario, all the ”mindfulness” samples are treated as ”avoidance,” resulting in a dataset with only two classes: acceptance and avoidance. The primary goal of this scenario is to investigate the performance of the five selected machine learning algorithms when classifying patients based on a simplified version of their pain coping strategies.

By merging the mindfulness class with the avoidance class, the focus shifts to evaluating the classifiers’ ability to distinguish between acceptance and avoidance coping strategies without the added complexity of the mindfulness category. The results of this classification process will be compared to the baseline performance established in the first scenario, which involved the original 3-class dataset. The performance metrics for each of the machine learning algorithms in this scenario are presented in Table 6.4 and Table 6.5, providing insights into the effect of reducing the number of classes on the classification outcomes.

	GBDT		ABDT		BDT		RF		ET	
	Value	SD	Value	SD	Value	SD	Value	SD	Value	SD
Accuracy	0.84	0.03	0.82	0.04	0.85	0.03	0.86	0.02	0.86	0.02
F1-score	0.56	0.08	0.57	0.08	0.57	0.09	0.57	0.09	0.55	0.08
Recall	0.97	0.03	0.93	0.04	0.97	0.03	0.99	0.02	0.99	0.02
Precision(PPV)	0.86	0.02	0.87	0.02	0.87	0.02	0.87	0.02	0.86	0.01
AUC	0.70	0.11	0.62	0.13	0.74	0.10	0.75	0.10	0.73	0.09
Specificity	0.15	0.12	0.20	0.13	0.15	0.13	0.14	0.12	0.11	0.10
NPV	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan

Table 6.4: Mindfullness Treated as Avoidance RFE

Upon comparing the results of Scenario 2 with the baseline performance established in Scenario 1, significant improvements can be observed in some metrics while others remain relatively unchanged or decline. The main points of comparison are as follows:

Accuracy: Scenario 2 shows substantial improvements in accuracy across all algorithms, with the highest value being 0.86 ± 0.02 compared to the previous maximum of 0.68 ± 0.04 in Scenario 1.

F1-score: The F1-score values have slightly improved in Scenario 2, with the highest F1-score now being 0.57 ± 0.09 compared to the previous maximum of 0.49 ± 0.08 in Scenario 1.

Recall: The recall values have also improved significantly in Scenario 2. The highest recall value is now 0.99 ± 0.02, compared to the previous maximum of 0.94 ± 0.04 in Scenario 1.

Precision (PPV): Precision values are relatively similar between the two scenarios, with the highest value in Scenario 2 being 0.87 ± 0.02 compared to the previous maximum of 0.69 ± 0.04 in Scenario 1.

Specificity: Specificity values in Scenario 2 have decreased for all algorithms, with the highest value being 0.20 ± 0.13 compared to the previous maximum of 0.37 ± 0.11 in Scenario 1.

	GBDT		ABDT		BDT		RF		ET	
	Value	SD	Value	SD	Value	SD	Value	SD	Value	SD
Accuracy	0.85	0.03	0.82	0.04	0.84	0.03	0.86	0.02	0.85	0.02
F1-score	0.57	0.10	0.55	0.07	0.57	0.09	0.57	0.08	0.57	0.08
Recall	0.97	0.03	0.94	0.04	0.96	0.03	0.99	0.02	0.98	0.02
Precision(PPV)	0.87	0.02	0.86	0.02	0.87	0.02	0.87	0.01	0.86	0.01
AUC	0.65	0.11	0.58	0.12	0.67	0.11	0.69	0.11	0.69	0.09
Specificity	0.16	0.14	0.15	0.11	0.16	0.13	0.14	0.11	0.13	0.10
NPV	nan	nan	0.32	0.24	nan	nan	nan	nan	nan	nan

Table 6.5: Mindfullness Treated As Avoidance SelectKBest

Upon applying the SelectKBest feature selection method to the same dataset as in the previous scenario, it was observed that the classification results remained relatively the same. This indicates that the selected machine learning algorithms' performance is consistent across different feature selection techniques when applied to this specific dataset.The similarity in results between the two feature selection methods suggests that the primary factors influencing classification performance in this scenario may be the inherent characteristics of the dataset rather than the feature selection method itself.

In conclusion, Scenario 2, which involves treating mindfulness samples as avoidance, demonstrates significant improvements in accuracy, F1-score, and recall for all algorithms. However, specificity values have declined.The trade-off between improved performance in some metrics and decreased performance in others should be taken into account.

It is important to note that the lower specificity observed in Scenario 2 might have significant implications in the context of detecting and preventing dysfunctional pain coping. A higher rate of false positives could lead to individuals with functional pain coping strategies (acceptance) being incorrectly identified as having dysfunctional pain coping strategies (avoidance), potentially resulting in unnecessary interventions or treatments. Therefore, striking a balance between sensitivity (recall) and specificity is essential to ensure that the classification model performs well in identifying both true positives and true negatives.

6.2.3 The Original Dataset with Mindfullness Treated as Acceptance

In the third scenario of the classification process, an alternative modification of the original dataset is explored. Contrary to the second scenario, this time, all the ”mindfulness” samples are treated as ”acceptance,” maintaining a two-class dataset: acceptance and avoidance. This scenario seeks to examine the performance of the selected machine learning algorithms when classifying patients based on this different simplification of their pain coping strategies, assuming that mindfulness practices are more closely aligned with acceptance. The classification results from this scenario are presented in Table 6.6 and Table 6.7 and will be compared to the outcomes of the previous scenarios to assess the impact of various class simplifications on classification performance.

	GBDT		ABDT		BDT		RF		ET	
	Value	SD	Value	SD	Value	SD	Value	SD	Value	SD
Accuracy	0.79	0.04	0.75	0.05	0.80	0.04	0.80	0.04	0.81	0.03
F1-score	0.64	0.08	0.62	0.08	0.65	0.08	0.64	0.09	0.63	0.09
Recall	0.32	0.13	0.35	0.14	0.33	0.14	0.29	0.14	0.27	0.13
Precision(PPV)	0.62	0.18	0.48	0.16	0.67	0.19	0.74	0.22	0.77	0.21
AUC	0.72	0.08	0.68	0.09	0.74	0.08	0.76	0.08	0.76	0.07
Specificity	0.93	0.04	0.88	0.05	0.95	0.04	0.97	0.03	0.97	0.02
NPV	0.82	0.03	0.81	0.03	0.82	0.03	0.81	0.03	0.81	0.03

Table 6.6: Mindfullness Treated as Acceptance RFE

	GBDT		ABDT		BDT		RF		ET	
	Value	SD	Value	SD	Value	SD	Value	SD	Value	SD
Accuracy	0.76	0.04	0.74	0.05	0.78	0.04	0.79	0.04	0.80	0.04
F1-score	0.60	0.08	0.59	0.08	0.64	0.08	0.64	0.08	0.64	0.09
Recall	0.26	0.12	0.28	0.13	0.34	0.14	0.31	0.13	0.31	0.14
Precision(PPV)	0.52	0.18	0.44	0.17	0.59	0.18	0.67	0.21	0.68	0.21
AUC	0.68	0.08	0.68	0.08	0.71	0.09	0.73	0.09	0.73	0.08
Specificity	0.92	0.04	0.89	0.05	0.92	0.05	0.94	0.04	0.95	0.04
NPV	0.80	0.03	0.80	0.03	0.82	0.03	0.82	0.03	0.82	0.03

Table 6.7: Mindfullness Treated as Acceptance SelectKBest

Comparing the results of Scenario 2 and Scenario 3, where mindfulness samples are treated as avoidance and acceptance respectively, we observe differences in the performance metrics. Scenario 2 shows higher accuracy, recall, and precision values than Scenario 3, suggesting that the algorithms perform better in identifying positive cases (avoidance) and avoiding false positives when mindfulness samples are treated as avoidance. On the other hand, Scenario 3 has higher F1-scores, AUC, specificity, and NPV values, indicating better performance in distinguishing between classes and predicting negative cases (acceptance).

Considering the performance metrics and the context that positive cases are ”avoidance” samples and negative cases are ”acceptance” samples, it appears that mindfulness samples are closer to avoidance than acceptance. This conclusion is based on the higher accuracy, recall, and precision values observed in Scenario 2, where mindfulness samples are treated as avoidance.

6.2.4 Original Dataset with Removed Mindfullness

In Scenario 4, the dataset used is the original dataset, but with mindfulness samples removed entirely, leaving only two classes: acceptance and avoidance. The need for this scenario stems from the desire to investigate the performance of the machine learning algorithms when focusing solely on the distinction between acceptance and avoidance, without the potential confounding factor of mindfulness samples.

By removing mindfulness samples from the dataset, this scenario aims to simplify the classification task and assess the efficacy of the algorithms in discriminating between the remaining two classes. The results obtained from this scenario as seen in Table 6.8 and Table 6.9 can be compared with those from the other scenarios to better understand the impact of mindfulness samples on the algorithms' performance and determine the most appropriate approach for analyzing the dataset.

	GBDT		ABDT		BDT		RF		ET	
	Value	SD	Value	SD	Value	SD	Value	SD	Value	SD
Accuracy	0.67	0.08	0.65	0.09	0.68	0.10	0.69	0.09	0.69	0.09
F1-score	0.63	0.09	0.61	0.10	0.66	0.10	0.66	0.10	0.65	0.11
Recall	0.77	0.11	0.73	0.14	0.77	0.12	0.80	0.11	0.82	0.11
Precision(PPV)	0.71	0.07	0.70	0.08	0.73	0.08	0.73	0.08	0.72	0.08
AUC	0.72	0.09	0.67	0.12	0.74	0.10	0.75	0.10	0.75	0.10
Specificity	0.50	0.15	0.51	0.17	0.55	0.16	0.52	0.16	0.48	0.17
NPV	0.60	0.14	0.56	0.15	0.62	0.15	0.65	0.17	0.64	0.17

Table 6.8: Removed Mindfullness RFE

	GBDT		ABDT		BDT		RF		ET	
	Value	SD	Value	SD	Value	SD	Value	SD	Value	SD
Accuracy	0.70	0.08	0.67	0.10	0.68	0.09	0.69	0.09	0.68	0.10
F1-score	0.67	0.09	0.64	0.11	0.66	0.10	0.66	0.10	0.64	0.11
Recall	0.81	0.11	0.75	0.13	0.76	0.13	0.79	0.11	0.80	0.12
Precision(PPV)	0.73	0.07	0.72	0.09	0.74	0.09	0.73	0.08	0.72	0.08
AUC	0.74	0.10	0.69	0.13	0.74	0.11	0.75	0.10	0.75	0.11
Specificity	0.52	0.16	0.54	0.17	0.56	0.18	0.53	0.17	0.50	0.17
NPV	0.66	0.15	0.59	0.16	0.62	0.15	0.63	0.16	0.63	0.18

Table 6.9: Removed Mindfullness SelectKBest

Given the different number of samples in each scenario, the comparison of the results becomes more context-dependent:

Scenario 2 shows higher accuracy, recall, and precision values than Scenario 4, which has a more balanced distribution of samples . This indicates that the models in Scenario 2 are better at identifying avoidance cases and avoiding false positives but may be biased towards the majority class (avoidance).

Scenario 3 , where mindfulness samples were treated as acceptance, has higher accuracy, specificity, and NPV values compared to Scenario 4. This suggests that the algorithms in Scenario 3 perform better in predicting acceptance cases and distinguishing between classes, but Scenario 4 is better at identifying avoidance cases (higher recall) and avoiding false positives (higher precision).

Given that both Scenario 2 and Scenario 4 are better at identifying avoidance cases than Scenario 3, it suggests that mindfulness samples may be closer to avoidance rather than acceptance. This observation could indicate that the nature of mindfulness samples has a stronger resemblance to the avoidance category in terms of the features being analyzed by the algorithms.

6.2.5 Optimised Dataset with Three Classes

In Scenario 5, we will examine the impact of using the optimized dataset derived through the data cleaning process described earlier. The optimized dataset has undergone a refinement based on specific criteria to remove invalid samples, such as those with a mean temperature below 28 degrees Celsius and samples with no relevant questionnaire responses. This dataset should provide a higher-quality and more representative sample of the data, which may lead to improved performance of the machine learning algorithms.

The need for testing Scenario 5 arises from the desire to determine whether the optimized dataset provides any significant benefits in the performance of the machine learning algorithms compared to the previous scenarios. By comparing the results of Scenario 5 to Scenario 1, we can evaluate the effectiveness of the data cleaning process and gauge the impact of using a more refined dataset on the overall performance and generalizability of the models. This analysis will help understand the importance of data quality in the context the experiment and the answers of the participants in the questionnaire.

Upon observing the reduction of the number of samples in the optimized dataset, it becomes evident that there is a significant decrease in the number of samples for both avoidance and acceptance, with the optimized dataset containing only 12 avoidance and 16 acceptance samples. The mindfullness samples in the optimized dataset have also decreased to 192. This reduction in sample sizes may raise concerns regarding the representativeness and statistical power of the dataset for training and testing the machine learning algorithms.Subsequently, Stratified k-fold cross-validation was not used in the testing of the models on the optimised dataset.

Interestingly, while it was expected that the majority of removed samples would be mindfullness samples, considering that participants who did not use any of the techniques taught in the lab experiment were to be excluded, the optimized dataset still maintains a substantial number of mindfullness samples. This observation could suggest that many participants might have engaged in the mindfulness technique, even if they did not explicitly report using avoidance or acceptance techniques in the questionnaire. Alternatively, it could also indicate potential inaccuracies or misinterpretations in the questionnaire responses, which may warrant further investigation and will be analysed in Chapter 7.

	GBDT		ABDT		BDT		RF		ET	
	Value	SD	Value	SD	Value	SD	Value	SD	Value	SD
Accuracy	0.89	0.06	0.85	0.05	0.90	0.05	0.92	0.05	0.92	0.04
F1-score	0.57	0.21	0.32	0.06	0.58	0.21	0.60	0.22	0.60	0.22
Recall	0.58	0.21	0.34	0.07	0.58	0.21	0.59	0.21	0.59	0.21
Precision(PPV)	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan
AUC	0.72	0.19	0.64	0.13	0.76	0.19	0.76	0.19	0.79	0.15
Specificity	0.79	0.10	0.67	0.04	0.80	0.10	0.80	0.10	0.80	0.10
NPV	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan

Table 6.10: Optimised Dataset with three classes RFE

	GBDT		ABDT		BDT		RF		ET	
	Value	SD	Value	SD	Value	SD	Value	SD	Value	SD
Accuracy	0.89	0.05	0.84	0.05	0.89	0.06	0.91	0.05	0.92	0.04
F1-score	0.58	0.21	0.31	0.04	0.58	0.22	0.59	0.22	0.59	0.22
Recall	0.59	0.21	0.33	0.06	0.59	0.22	0.59	0.21	0.59	0.21
Precision(PPV)	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan
AUC	0.78	0.17	0.71	0.13	0.78	0.17	0.80	0.16	0.84	0.13
Specificity	0.79	0.10	0.66	0.03	0.79	0.11	0.79	0.10	0.79	0.10
NPV	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan

Table 6.11: Optimised Dataset with three classes SelectKBest

Scenario 5 demonstrates a significant improvement in the performance of the machine learning algorithms compared to Scenario 1 as seen from Table 6.10 and Table 6.11. Specifically, the highest accuracy value in Scenario 5 is 0.92 ± 0.04, a notable increase from Scenario 1's highest accuracy value of 0.68 ± 0.04. Furthermore, the highest specificity value in Scenario 5 is 0.80 ± 0.10, a substantial improvement over the highest value of 0.37 ± 0.12 in Scenario 1. Lastly, the highest AUC value in Scenario 5 is 0.79 ± 0.15, which is slightly higher than the highest value of 0.74 ± 0.06 in Scenario 1.

These improvements suggest that the data cleaning process has led to a more accurate and representative dataset that enables better distinction between the classes and more reliable negative case predictions. However, the decrease in recall values indicates a reduced ability of the algorithms in Scenario 5 to identify positive cases.

It is important to note that the significant reduction in the number of avoidance and acceptance samples in the optimized dataset may have an impact on the performance and generalizability of the algorithms. The limited sample size could be a potential factor contributing to the lower recall values observed in Scenario 5.

Additionally, the varying F1-score values across different models in Scenario 5 suggest that it would be important to carefully select the best model based on specific use cases and the desired balance between precision and recall for each class, including mindfulness.

6.2.6 Optimised Dataset with Mindfullness Treated as Avoidance

In this scenario, we aim to assess the performance of machine learning algorithms when trained on the optimized dataset with the original three classes, but all mindfulness samples are treated as avoidance samples. This results in a dataset with only two classes, acceptance and avoidance. The need for testing this scenario arises from the hypothesis from previous scenarios that mindfulness samples may be closer to avoidance rather than acceptance, and by combining these two classes, we can potentially improve the classification results. This scenario also allows us to compare the algorithms' performance with previous scenarios, providing insights into the impact of merging mindfulness and avoidance samples on classification results in the context of the optimized dataset.

	GBDT		ABDT		BDT		RF		ET	
	Value	SD	Value	SD	Value	SD	Value	SD	Value	SD
Accuracy	0.95	0.04	0.95	0.04	0.95	0.03	0.96	0.03	0.96	0.03
F1-score	0.73	0.21	0.75	0.22	0.73	0.21	0.75	0.22	0.75	0.22
Recall	0.99	0.03	0.99	0.03	0.99	0.02	1.00	0.01	1.00	0.00
Precision(PPV)	0.96	0.03	0.96	0.03	0.96	0.03	0.96	0.03	0.96	0.03
AUC	0.84	0.23	0.83	0.22	0.84	0.19	0.85	0.19	0.83	0.21
Specificity	0.50	0.43	0.52	0.44	0.47	0.42	0.47	0.42	0.48	0.42
NPV	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan

Table 6.12: Optimised Dataset with Mindfullness Treated as Avoidance RFE

	GBDT		ABDT		BDT		RF		ET	
	Value	SD	Value	SD	Value	SD	Value	SD	Value	SD
Accuracy	0.95	0.04	0.92	0.05	0.95	0.04	0.96	0.03	0.96	0.03
F1-score	0.73	0.21	0.68	0.18	0.73	0.21	0.75	0.22	0.75	0.22
Recall	0.99	0.02	0.96	0.04	0.98	0.03	1.00	0.01	1.00	0.01
Precision(PPV)	0.96	0.03	0.96	0.03	0.96	0.03	0.96	0.03	0.96	0.03
AUC	0.83	0.18	0.82	0.20	0.80	0.22	0.82	0.19	0.85	0.17
Specificity	0.48	0.42	0.48	0.42	0.48	0.42	0.48	0.42	0.48	0.42
NPV	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan

Table 6.13: Optimised Dataset with Mindfullness Treated as Avoidance dataset SelectKBest

In conclusion, Scenario 6 demonstrates a significant improvement in the performance of the machine learning algorithms compared to Scenario 2, as seen from Table 6.12 and Table 6.13. Specifically, the highest accuracy increased from 0.86 ± 0.02 in Scenario 2 to 0.96 ± 0.03 in Scenario 6, and the highest F1-score improved from 0.57 ± 0.09 to 0.75 ± 0.22. Additionally, the highest precision (PPV) increased from 0.87 ± 0.02 in Scenario 2 to 0.96 ± 0.03 in Scenario 6.

One of the most notable improvements is the increase in specificity. In Scenario 6, the highest specificity value is 0.52 ± 0.44, compared to 0.20 ± 0.13 in Scenario 2. This implies that, in Scenario 6, the algorithms can accurately classify ”acceptance” samples roughly half the time (0.50), compared to only 20 % of the time in Scenario 2. This significant increase in specificity indicates a better ability of the algorithms to correctly identify true negatives and reduce false positives, which is crucial in real-world applications where false alarms could lead to wasted resources or unnecessary interventions.

Despite the reduction in the number of samples for each class in Scenario 6 (16 acceptance and 204 avoidance samples) compared to Scenario 2 (79 acceptance and 440 avoidance samples), the algorithm’s ability to classify instances and predict outcomes has improved. This improvement may be attributed to the data cleaning process, which resulted in a more representative and accurate dataset.

6.2.7 Optimised Dataset with Mindfullness Treated as Acceptance

Similarly to the previous scenario, Scenario 7 evaluates the performance of machine learning algorithms when trained on the optimized dataset. However, in this case, all mindfulness samples are treated as acceptance samples, resulting in a dataset with only two classes: acceptance and avoidance. The main motivation for testing this scenario is to investigate whether the optimized dataset demonstrates improved performance compared to the original dataset in Scenario 3, where mindfulness samples were also treated as acceptance samples. By comparing the results of Scenario 7 with those of Scenario 3, we can gain insights into the impact of data optimization on the classification outcomes when mindfulness and acceptance samples are combined.

	GBDT		ABDT		BDT		RF		ET	
	Value	SD	Value	SD	Value	SD	Value	SD	Value	SD
Accuracy	0.94	0.03	0.94	0.04	0.95	0.03	0.96	0.02	0.96	0.02
F1-score	0.62	0.20	0.62	0.19	0.64	0.21	0.65	0.22	0.65	0.22
Recall	0.30	0.42	0.32	0.43	0.30	0.42	0.30	0.42	0.30	0.42
Precision(PPV)	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan
AUC	0.68	0.32	0.59	0.32	0.72	0.29	0.70	0.31	0.69	0.27
Specificity	0.98	0.03	0.98	0.03	0.99	0.03	1.00	0.00	1.00	0.00
NPV	0.96	0.03	0.96	0.03	0.96	0.02	0.96	0.02	0.96	0.02

Table 6.14: Optimised Dataset with Mindfullness Treated as Acceptance RFE

	GBDT		ABDT		BDT		RF		ET	
	Value	SD	Value	SD	Value	SD	Value	SD	Value	SD
Accuracy	0.94	0.04	0.94	0.04	0.94	0.04	0.95	0.03	0.96	0.03
F1-score	0.65	0.21	0.64	0.21	0.62	0.20	0.64	0.21	0.65	0.22
Recall	0.38	0.46	0.35	0.45	0.30	0.43	0.30	0.42	0.30	0.42
Precision(PPV)	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan
AUC	0.72	0.29	0.74	0.27	0.74	0.27	0.75	0.27	0.73	0.28
Specificity	0.97	0.04	0.97	0.03	0.98	0.03	0.99	0.02	1.00	0.01
NPV	0.97	0.03	0.96	0.03	0.96	0.03	0.96	0.02	0.96	0.02

Table 6.15: Optimised Dataset with Mindfullness Treated as Acceptance SelectKBest

In conclusion, based on the comparison between Scenarios 3 and 7, the optimized dataset in Scenario 7 demonstrates improved performance, as seen in Table 6.14 and table 6.15 in terms of accuracy (0.96 ± 0.02 vs. 0.81 ± 0.03), specificity (1.00 ± 0.00 vs. 0.97 ± 0.02), and NPV (0.96 ± 0.02 vs. 0.82 ± 0.03). However, Scenario 3 outperforms Scenario 7 in F1-score (0.65 ± 0.08 vs. 0.65 ± 0.22) and AUC values (0.76 ± 0.07 vs. 0.72 ± 0.29). Recall values are generally similar (0.35 ± 0.14 for Scenario 3 and 0.32 ± 0.43 for Scenario 7), but with much higher standard deviations in Scenario 7.

Although there are some improvements in Scenario 7, the differences are not substantial, and the results also reveal trade-offs between different performance metrics. In both Scenarios 3 and 7, where mindfulness samples are treated as acceptance samples, it appears that the algorithms perform similarly, suggesting that mindfulness samples are closer to avoidance samples rather than acceptance samples.

6.3 Comparison of Results with Previous Work

This section compares the results of the current work with the four previous related analyses that were described in Section 3.2. Table 6.16 shows the candidate features and the selected features in each analysis, as well as the selected classifier in each one.

In the first three studies the most effective features came from ECG and COR signals. In [6] SCR peak that came from GSR gave the best results. In this study, features that came

Table 6.16: Results of previous analyses and current work

	Selected Features	Best Classifier Overall	Average Accuracy of the Best Classifier (%)
Galazis, 2017 [4]	Smoking: GSR	RF	66
	Eating Disorder: ECG, BI-AAQ (questionnaire)		
	Anxiety: COR		
Trigeorgi, 2018 [12]	Smoking: ECG_rmssd	RF	73
	Eating Disorder: ECG_sdnn, COR_mean		
	Anxiety: ECG_bpm, ECG_rmssd, ECG_pnn20, COR_mean		
Demosthenous,2019 [5]	All: GSR_mean, COR_mean	BDT (RWM)	90
E.Georgiou,2022 [6]	Pain: SCRwatch_meanPeakAmp, ECG_heartrate	BDT (RWM)	85
Current Study	Acceptance vs Avoidance 'hrv_rmssd', 'hrv_sdsd', 'hrv_sdnn', 'hrv_lf', 'hrv_hf', 'temp_mean', 'bpm', 'ibi'	RF	69
	Mindfullness as Avoidance 'hrv_pnni_50', 'hrv_pnni_20', 'hrv_hf', 'temp_mean', 'bpm', 'ibi', 'pnn20', 'pnn50'	RF	86
	Mindfullness as Acceptance : 'hrv_mean_hr', 'hrv_mean_nni', 'hrv_rmssd', 'hrv_sdsd', 'hrv_sdnn', 'hrv_lf', 'hrv_hf', 'lf/hf'	ET	80

from PPG sensor and therefore HRV features were the most common across all cases and experiments.

As regards to the most effective Machine Learning Algoritm, Random Forest performed the best in the first two studies, while Bagging Decision Tree was the best in the last two.In the current study, Extra Tress and Random Forest perform the best across different cases, with the other algorithms not significantly behinf.It is worth notting that even though in some cases, higher accuracy was achieved (around 95% when using the optimised dataset), we could not consider it as valid as previous works because of data imbalance and small sample numbers in general.Nevertheless, the accuracy achieved is very satisfactory.

Chapter 7

Discussion

Contents

7.1 Summary

This thesis is the continuation of a series of experiments and analyses of emotional coping using psychophysiological features. It focuses on the Functional Versus Dysfunctional Coping in Real Time experiment, the first work that utilised real-time data instead of data extracted from in-lab experiments. The data were acquired from an experiment conducted by the Department of Psychology of the University of Cyprus.The experiment required participants to wear the Empatica E4 wearable device for 3 days and were prompted to questions on an app pre-installed by the researchers on the provided smartphones.The signals that were recorded were Photoplethysmography (PPG), Electrodermal Activity (EDA), Accelerometer (ACC) and Temperature (TEMP).

Due to the fact that the dataset contained recordings spanning 3 days for each participant, careful data extraction was needed.More specifically, time frames of 10 minutes were extracted from the moment the patients answered their questionnaire , and each one of those was treated as a different sample, resulting in a much larger dataset than previous studies.Afterwards, for each raw signal, multiple features were extracted.

The selection of the most relevant features was a key focus in this thesis, as it plays a crucial role in the performance and interpretability of machine learning models. Throughout the thesis, different feature selection approaches were discussed, including Wrapper, Filter, and Embedded methods. These methods were used to identify the most relevant features from various psychophysiological signals such as PPG and EDA. The selected features were then used to train and evaluate different machine learning models, such as Adaptive Boosting, Gradient Boosting Decision Tree, Bagging Decision Tree, Random Forest, and Extra Trees, in order to select the best-performing model for the given task.

Emphasis was given in the categorisation of the mindfullness samples into one of the two classes, acceptance and avoidance. By applying the afforementioned models on different versions of the original dataset, were mindfullness was treated as avoidance or as acceptance, many conclusions could be made. Most importantly, it was shown that those samples are indeed closer to avoidance rather than acceptance, giving an average accuracy of 86

7.2 Future Work

The present study, while offering promising results, also reveals areas for future improvement and expansion, particularly in data collection, feature selection, algorithm application, and questionnaire design. These enhancements could potentially lead to even more accurate and robust findings.

Firstly, the current dataset's imbalance potentially biases the classifiers towards the majority class. Future research could involve the collection of a larger, more balanced dataset to provide more accurate outcomes. Because of the nature of the experiment and its difficulty in collecting more samples in a small time window, alternatively, synthetic minority over-sampling techniques, which were used in previous work [5], could be employed to generate artificial samples from the minority class and alleviate the class imbalance issue.

In addition, it was observed that many samples contained unrealistic values, such as temperatures below 28 degrees Celsius, suggesting that patients were not wearing the device during the questionnaire. Future studies could implement stricter protocols or use additional sensors to confirm whether patients are indeed wearing the devices during the experiment.

Moreover, many questions in the current questionnaire showed little relevance to the pain coping strategies employed by participants. Future iterations could focus on refining and validating these questionnaires to enhance their reliability and predictive power. For instance, leveraging qualitative research methods to understand patients' experiences better could inform the creation of more relevant and insightful questions.

As regards to the responses, the most critical question used to categorise patients into pain coping strategies were sometimes unclear, possibly leading to misclassification. Future work could involve refining the phrasing of these questions or providing additional guidance to participants to ensure more accurate responses.

Lastly, in terms of machine learning algorithms, future work could investigate more sophisticated models, such as ensemble methods, deep learning models, or Neural Networks. These models could potentially offer improved performance by capturing more complex patterns within the data.

Bibliography

[1] Brandon Farnsworth. What is ecg and how does it work? 8 2021. [Online]. Accessed on 7 May 2022.

[2] David Castaneda, Andres Esparza, Mohammad Ghamari, Cinthia Soltanpur, and Homer Nazeran. A review on wearable photoplethysmography sensors and their potential future applications in health care. *Int J Biosens Bioelectron*, 4(4):195–202, 2018.

[3] John Wilson. What is facial emg and how does it work? *iMotions*, 2018. [Online]. Accessed on 7 May 2022.

[4] C.Galazis. Non-Intrusive Physiological Wearable Devices for Identifying Individual Difference Parameters Using Supervised Classification Learning Algorithms. 2017.

[5] G. Demosthenous. Machine Learning Approach to Predict Emotional Coping Using Psychophysiological Signals, MSc thesis, Department of Computer Science, University of Cyprus. 2019.

[6] E.Georgiou. Feature Selection and Training of Machine Learning Algorithms to Classify Functional versus Dysfunctional Coping with Acute Pai, Department of Computer Science, University of Cyprus. 2022.

[7] A. Trigiorgi. Μελέτη Μεθόδων Μηχανικής και Βαθιάς Μάθησης και Εφαρμογή σε Ψυχομετρικά Δεδομένα, Diploma Project, Department of Computer Science, University of Cyprus. 2016.

[8] E. Morgan. All about hrv part 2: Interbeat intervals and time domain stats. *MindWare*, 2017.

[9] C. E. Ackerman. How Does Acceptance And Commitment Therapy (ACT) Work?," 29 March 2022. [Online]. Available: https://positivepsychology.com/act-acceptanceand-commitment-therapy/. 2022.

[10] Yasir Sattar and Lovy Chhabra. Electrocardiogram. *StatPearls Publishing*, 2021.

[11] Euan Ashley and Josef Niebauer. Cardiology explained. 2004.

[12] Ά. Τριγιώργη. Εξόρυξη Γνώσης από Ψυχοφυσιολογικά Δεδομένα και Συγκριτική Αξιολόγηση Αλγορίθμων Μηχανικής Μάθησης. 2018.

[13] iMotions. Galvanic skin response (gsr): The complete pocket guide. 2020. [Online]. Accessed on 7 May 2022.

[14] Andrej A. Romanovsky. Skin temperature: its role in thermoregulation. *Acta Physiol (Oxf)*, 2014.

[15] R. Brooks and K. Dahlke. Understanding the 3 categories of machine learning – ai vs. machine learning vs. data mining 101 (part 2)," 17 october 2017. [online]. available: https://www.guavus.com/ai-vs-machine-learing-vs-data-mining-whats-big-difference-part-2/. 23 May 2022.

[16] V. Kurama. Gradient boosting in classification: Not a black box anymore!," 2020. [online]. available: https://blog.paperspace.com/gradient-boosting-for-classification/. 18 April 2022.

[17] Sunil. Quick introduction to boosting algorithms in machine learning. 11 2015. [Online]. Accessed on 16 April 2022.

[18] Vamsi Kurama. Gradient boosting in classification: Not a black box anymore! 2020. [Online]. Accessed on 18 April 2022.

[19] Jason Brownlee. A gentle introduction to ensemble learning algorithms. 4 2021. [Online]. Accessed on 16 April 2022.

[20] Sumedha Eswar R. Understanding random forest. 6 2021. [Online]. Accessed on 16 April 2022.

[21] Pablo Aznar. What is the difference between extra trees and random forest? 6 2020. [Online]. Accessed on 18 April 2022.

[22] Jason Brownlee. A gentle introduction to k-fold cross-validation. 5 2018. [Online]. Accessed on 21 April 2022.

[23] Rohit Toshniwal. How to select performance metrics for classification models. 1 2020. [Online]. Accessed on 10 March 2023.

[24] E4 wristband, the legacy industry-standard for real-world, and academic research.

[25] Empatica. Empatica e4 specifications. [Online]. Accessed in 2022.

[26] E4 wristband.

[27] E4 data.

[28] E. K. Sandoz, K. G. Wilson, R. M. Merwin, and K. K. Kellum. Assessment of body image flexibility: The body image-acceptance and action questionnaire. *Journal of Contextual Behavioral Science*, 2(1-2):39–48, 2013.

[29] P. Konstantinou, A. Trigeorgi, C. Georgiou, A. T. Gloster, G. Panayiotou, and M. Karekla. Functional versus dysfunctional coping with acute pain: An experimental comparison of acceptance vs. avoidance coping.

[30] Saul Shiffman, Arthur A Stone, and Michael R Hufford. Ecological momentary assessment. *Annu. Rev. Clin. Psychol.*, 4:1–32, 2008.

[31] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification And Regression Trees*. New York, 2017.

[32] David C. Swanson. Signal processing for intelligent sensor systems with matlab. 2011.

[33] Maria Karekla, Giorgos Demosthenous, Chryssis Georgiou, Pinelopi Konstantinou, Andria Trigiorgi, Maria Koushiou, Georgia Panayiotou, and Andrew T. Gloster. Machine learning advances the classification and prediction of responding from psychophysiological reactions. *Journal of Contextual Behavioral Science (JCBS)*, 26:36–43, October 2022.

[34] Institute for Dynamic Systems and Control, ETH Zurich. FLIRT: Fast Localized Intersection-based Road Traffic Prediction, 2023.

[35] J. Shukla, M. Barreda-Ángeles, J. Oliver, G. C. Nandi, and D. Puig. Feature extraction and selection for emotion recognition from electrodermal activity. *IEEE Transactions on Affective Computing*, 12(4):857–869, 2021.

[36] Christopher Glen Thompson, Rae Seon Kim, Ariel M. Aloe, and Betsy Jane Becker. Extracting the variance inflation factor and other multicollinearity diagnostics from typical regression results. *Basic and Applied Social Psychology*, 39(2):81–90, 2017.

[37] Vaishali Verma. A comprehensive guide to feature selection using wrapper methods in python. 10 2020. [Online]. Accessed on 12 April 2022.