Individual Diploma Thesis

MACHINE LEARNING CLASSIFICATION OF PSYCHOPHYSIOLOGICAL DATA FROM CANCER PATIENTS IN LABORATORY AND REAL TIME SETTINGS USING WAREABLE SENSORS

Giorgos Sofroniou

University Of Cyprus

Department of Computer Science

May 2025

University of Cyprus

Department of Computer Science

Machine Learning Classification of Psychophysiological Data from Cancer Patients in Laboratory and Real Time Settings using Wareable Sensors

Giorgos Sofroniou

Supervisor
Chryssis Georgiou

The Individual Diploma Thesis was submitted in partial fulfillment of the requirements for the acquisition of the Informatics degree from the Department of Computer Science of the University of Cyprus.

May 2025

Acknowledgement

I would like to express my sincere gratitude to my supervisor, Dr. Chryssis Georgiou, Professor at the Department of Computer Science of the University of Cyprus, for trusting me without hesitation and giving me the opportunity to undertake and complete this topic. I would also like to thank him for the invaluable help and support he offered me throughout all these months during the execution of my thesis project.

I would also like to thank PhD student Andria Trigeorgi for her continuous supervision and invaluable assistance throughout the thesis project. Her constant support and the feedback she provided whenever a second opinion was needed were truly essential for the successful completion of the project.

Furthermore, I would like to thank the Department of Psychology at the University of Cyprus for conducting the experiments and providing us with the patient data samples.

Finally, I would like to express my heartfelt gratitude to my family, my friends and my fellow students for their continuous support and for always being by my side whenever I needed them. Without their presence and encouragement, the successful completion of this thesis project would not have been possible.

Abstract

In recent years, the integration of Artificial Intelligence (AI) and wearable devices has led to significant progress in healthcare, including Acceptance and Commitment Therapy (ACT), a psychotherapy approach that helps individuals accept difficult emotions and act in line with their values. This thesis explores whether wearable devices can reliably detect functional and dysfunctional pain coping strategies, both in laboratory settings and in real-life situations.

Unlike previous studies focused on university students and conditions such as anxiety and eating disorders, this research involves two experiments conducted on cancer patients undergoing treatment. The first experiment, was an emotional imagery task conducted by the Department of Psychology at the University of Cyprus, required participants to complete six guided imagery trials following a five-minute heart rate variability (HRV) assessment. Psychophysiological data were collected using Shimmer wearable sensors, along with responses to psychological questionnaires. Focus was placed on the DASS-S (stress level) and AAQ-II (psychological flexibility) scores, which were used for labeling participants based on literature-recommended and median thresholds, resulting in four scenarios.

For the second experiment, participants wore Empatica E4 wristbands and answered ecological momentary assessment (EMA) questions through a mobile app regarding their emotional state, pain, and social context. Both experiments recorded Photoplethysmography (PPG) and Electrodermal Activity (EDA) signals, while the second also included accelerometer (ACC) and temperature (TEMP) data.

After signal extraction, relevant features were derived and filtered using three established feature selection methods. Four supervised machine-learning models, Adaptive Boosting, Gradient Boosting, Random Forest, and Extra Trees were trained on the selected features. Oversampling techniques were applied to address class imbalance where necessary.

The results showed that in the emotional imagery experiment, the Extra Trees algorithm performed best in most scenarios, while Bagging Decision Tree led in one. Random Oversampler consistently proved to be the most effective oversampling method. In the EMA experiment, the combination of Extra Trees and Random Oversampler achieved the highest classification performance, demonstrating the potential of AI and wearable technology to support psychological interventions for cancer patients.

The study confirmed that psychophysiological signals from wearable devices can reliably classify coping strategies, with model accuracy exceeding 80% in several cases. These findings support the integration of AI into ACT-based interventions and highlight the importance of selecting relevant features and interpreting signals based on context, reinforcing the value of personalized, real-time monitoring in clinical decision-making.

Contents

Chapter 1 - Introduction

1.1 Motivation

In recent years, wearable technologies like smartwatches and fitness bands have seen widespread adoption. These devices are capable of tracking various psychophysiological signals, including heart rate and skin conductance. Notable examples include Electrocardiogram (ECG) [1], Electrodermal Activity (EDA) [2], and facial electromyography (fEMG) [3].

Several studies have explored the analysis of such signals (e.g., [4],[5],[6],[7],[8]) with most relying on data collected from non-portable, stationary equipment. Only a few, such as [6] and [8], have utilized data from wearable devices. Additionally, previous research has primarily focused on heart rate variability (HRV) time-domain features derived from ECG signals [9]. This study seeks to expand on that by incorporating features extracted from EDA and Photoplethysmography (PPG) signals, using data obtained from existing patients rather than university students, as was common in earlier work. The results of this research may offer valuable insights for healthcare applications and contribute to improving patient care and intervention strategies.

1.2 Goals of Study

This thesis utilizes data collected from two experiments on pain management techniques, conducted by the Department of Psychology at the University of Cyprus. The primary objective of this study is to contribute to the integration of *Acceptance and Commitment Therapy* (ACT) [51] into the daily lives of real cancer patients. ACT is a psychotherapy approach that encourages individuals to engage with their thoughts and emotions rather than avoiding or suppressing them. It has been shown to be effective in treating conditions such as obsessive-compulsive disorder (OCD), anxiety, and depression.

Within the framework of ACT, individuals are categorized into two groups based on their reactions to emotional experiences: the *functional* group, comprising individuals who accept their internal experiences and cope with them effectively, characterized by low stress and high psychological flexibility and the *dysfunctional* group, consisting of individuals who engage in avoidance strategies, leading to high stress and low psychological flexibility. It is important to note that a person's classification can vary depending on their environment and circumstances.

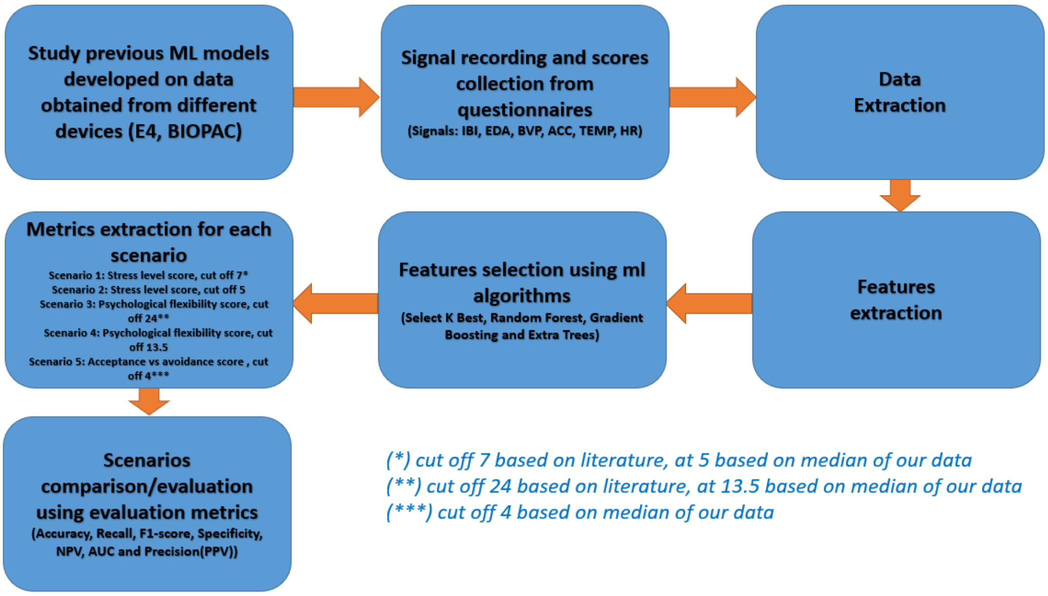
The primary goal of this thesis is to develop machine-learning models capable of accurately classifying individuals into functional or dysfunctional categories based on their pain coping strategies. In the emotional imagery experiment, the functional group includes patients with low stress or high psychological flexibility, while the dysfunctional group includes those with high stress or low psychological flexibility. In the EMA experiment, classification is based on whether the patient belongs to the "acceptance" (functional) or "avoidance" (dysfunctional) group.

A key focus of this research is to investigate whether psychophysiological signals recorded from wearable devices are sufficient for training effective machine learning algorithms. Additionally, the study examines feature selection techniques from three major categories, aiming to compare their performance and identify the most informative subset of features.

Ultimately, both experiments seek to accurately distinguish between functional and dysfunctional coping strategies related to pain and emotional distress. By combining these analyses, this thesis aspires to enhance the understanding of psychophysiological correlations of psychological flexibility, providing valuable insights for improving patient care and therapeutic interventions.

1.3 Methodology

The outline of the methodology used is shown in Figure 1.1.


Study previous ML models developed on data obtained from different devices (E4, BIOPAC)
Signal recording and scores collection from questionnaires
(Signals: IBI, EDA, BVP, ACC, TEMP, HR)
Data Extraction
Features extraction
Features selection using ml algorithms
(Select K Best, Random Forest, Gradient Boosting and Extra Trees)
Metrics extraction for each scenario
Scenario 1: Stress level score, cut off 7*
Scenario 2: Stress level score, cut off 5
Scenario 3: Psychological flexibility score, cut off 24**
Scenario 4: Psychological flexibility score, cut off 13.5
Scenario 5: Acceptance vs avoidance score , cut off 4***
Scenarios comparison/evaluation using evaluation metrics
(Accuracy, Recall, F1-score, Specificity, NPV, AUC and Precision(PPV))

(*) cut off 7 based on literature, at 5 based on median of our data
(**) cut off 24 based on literature, at 13.5 based on median of our data
(***) cut off 4 based on median of our data

Figure 1.1: An outline of the methodology used

We began this thesis by reviewing past implementations of machine learning applied to psychophysiological data collected from wearable devices such as the Empatica E4 and BIOPAC, commonly used in the related studies [8],[10],[11] and [12].

Subsequently, the Department of Psychology at the University of Cyprus conducted two experiments using Shimmer [31] and Empatica E4 [32] devices to collect physiological signals. The emotional imagery experiment was conducted in a controlled laboratory setting, whereas the EMA study collected data in real-life conditions. In both cases, Photoplethysmography (PPG) and Electrodermal Activity (EDA) signals were obtained. The EMA experiment further included accelerometer (ACC) and temperature (TEMP) data, offering a broader scope of physiological monitoring.

In addition to signal data, each participant completed self-report psychological questionnaires. For the emotional imagery experiment, the Depression Anxiety Stress Scale – Stress subscale (DASS-S) and the Acceptance and Action Questionnaire-II (AAQ-II) were used to assess perceived stress and psychological flexibility, respectively. For the EMA dataset, binary labels representing acceptance or avoidance were generated based on participants' responses to a specific self-report item. Prior to model training, standard preprocessing and signal-cleaning techniques were applied using Python [13], followed by extraction of relevant time-domain and statistical features.

To reduce feature dimensionality and enhance learning performance, four feature selection techniques were used: Random Forest feature importance, SelectKBest, Gradient Boosting Decision Tree, and Extra Trees [13]. The classification process was structured around five scenarios—four from the emotional imagery experiment and one from the EMA phase, each reflecting different thresholds or scoring interpretations:

- Scenario 1: DASS stress score, threshold = 7 (literature-based)
- Scenario 2: DASS stress score, threshold = median
- Scenario 3: AAQ-II psychological flexibility score, threshold = 24 (literature-based)
- Scenario 4: AAQ-II psychological flexibility score, threshold = median
- Scenario 5: EMA-based classification into acceptance or avoidance, threshold = 4 (median)

Each scenario was evaluated using five supervised binary classification algorithms: Adaptive Boosting, Gradient Boosting Decision Tree, Bagging Decision Tree, Random

Forest, and Extra Trees [13]. Following the initial evaluation, oversampling methods were applied to manage class imbalance and further assess classification performance under more balanced data distributions.

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's short description
2	Provides an overview of the background knowledge upon which this thesis is based. It first explains the psychophysiological signals recorded in the Psychology lab, followed by an analysis of the machine learning algorithms, methodologies, and evaluation metrics used. Finally, it briefly describes the devices employed throughout the experiments.
3	Overviews the four previous experiments and presents a detailed explanation of the current experimental work.
4	Describes the complete process for obtaining the features used to train the algorithms, starting from the raw signal recordings. It outlines the methodology applied to extract samples from patients for each experiment and details the features that were subsequently extracted.
5	Analyzes the feature selection methods applied and presents the final subset of selected features for each dataset.
6	Presents both the original classification results and the improved results obtained after applying oversampling techniques.
7	Summarizes the overall work conducted in this thesis and proposes directions for future research and improvements.

Table 1.1 Document organization

Chapter 2 - Background Knowledge

2.1 Psychophysiological Signals

The Department of Psychology at the University of Cyprus conducted a series of five experiments, mentioned in Chapter 3. During the experiment under investigation, the psychophysiological signals discussed in this section were collected.

2.1.1 Electrocardiogram (ECG)

ECG is a method used to detect the heart’s electrical behavior without penetrating the body [14]. It records waveforms from the skin’s surface, showing common elements known as P, QRS, and T waves [15] (see Figure 2.1). Signals gathered through ECG allow the study of several types of features, including time-related, frequency-related, and spectral ones. This work prioritizes time-based analysis, which has shown to be more aligned with our aims, supported by earlier publications [16]. Time-based data often centers around Heart Rate Variability (HRV), which tracks how the timing between heartbeats changes. This is typically measured through RR intervals, the span between two R points in the signal, with the R wave located within the QRS region.

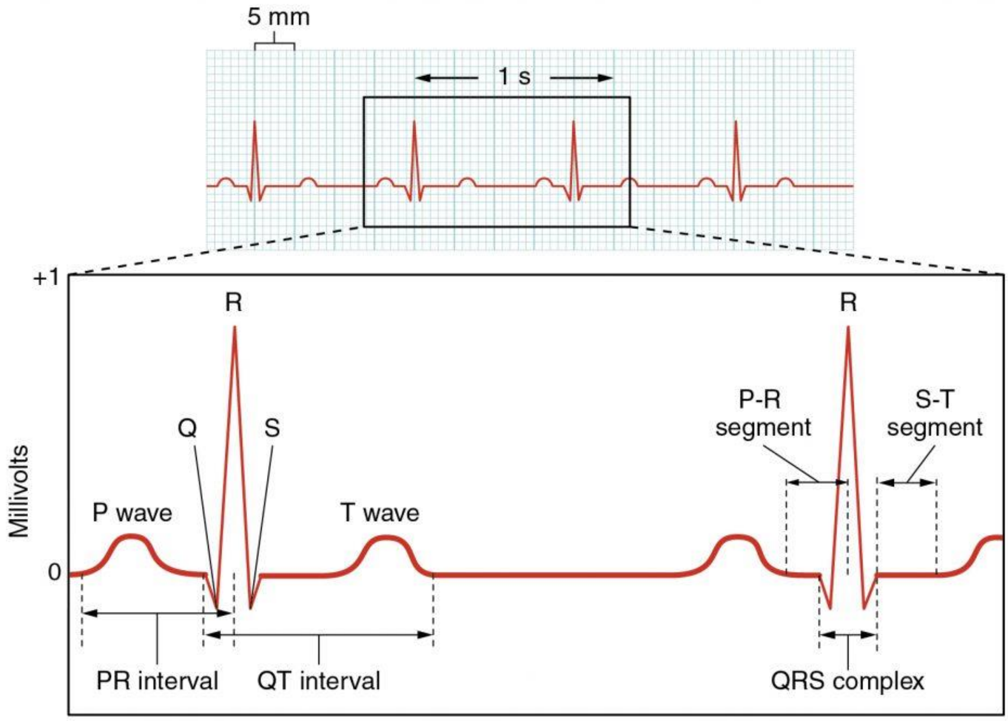


Figure 2.1: Graphical representation of the ECG signal

2.1.2 Photoplethysmography (PPG)

Photoplethysmography (PPG) is a widely used, non-invasive technique that detects volumetric changes in blood circulation within the peripheral microvasculature by means of light-based sensing methods [1]. Typically, a PPG sensor consists of an LED that emits light—commonly green or infrared—and a photodetector that captures the reflected or transmitted light through the skin. These light fluctuations arise due to variations in blood volume occurring with each heartbeat[1][2] as seen in Figure 2.2. By analyzing the resulting optical signal, researchers can derive important cardiovascular indicators, such as heart rate and heart rate variability (HRV). Specifically, HRV is determined from the intervals between consecutive peaks in the PPG waveform, which correspond to the time difference between successive heartbeats and are analogous to the RR intervals identified in electrocardiogram (ECG) analysis [2]. Section 4.2.1 provides an in-depth description of the features that can be derived from heart rate variability.

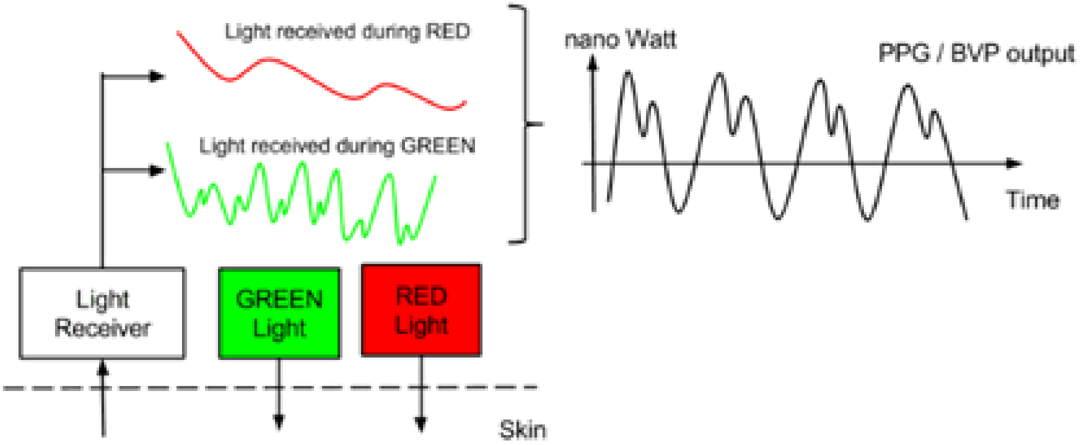


Figure 2.2: Graphical Representation of PPG Signal

2.1.3 Electrodermal Activity (EDA)

Electrodermal Activity (EDA), also referred to as Galvanic Skin Response (GSR) [17] or Skin Conductance (SC), describes variations in the skin's electrical properties resulting from sweat gland activity. These changes are closely associated with a person's emotional state or level of physiological arousal. EDA can be measured by applying a small electrical potential between two points on the skin and recording the resulting current flow. This signal provides valuable insights, particularly in the assessment of pain, and has a wide range of applications in clinical and psychological contexts. For a more comprehensive explanation of EDA signal formation and the features that can be extracted from it, see Section 4.2.2.

2.1.4 Inter-Beat Interval (IBI)

The Inter-Beat Interval (IBI) [18], illustrated in Figure 2.3, represents the time interval between two consecutive heartbeats. This metric serves as the basis for calculating the instantaneous heart rate.

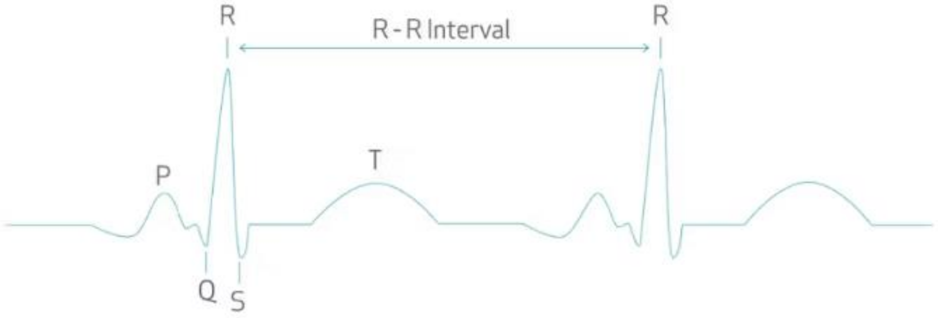


Figure 2.3: Inter-Beat Interval

2.1.5 Blood Volume Pulse (BVP)

The Photoplethysmography (PPG) sensor primarily outputs the Blood Volume Pulse (BVP) signal, which serves as a key indicator of cardiovascular activity [19]. This signal is computed using a proprietary method that integrates the optical data collected under green and red LED illumination, as depicted in Figure 2.2. The resulting BVP waveform is recorded at a uniform sampling frequency of 64 Hz, corresponding to 64 data points per second.

2.1.6 Acceleration (ACC)

The 3-axis accelerometer is an advanced motion-sensing component capable of detecting acceleration forces independently along the X, Y, and Z directions. By capturing variations in acceleration, it provides valuable data related to body posture, movement patterns, and exertion levels. This information is instrumental in recognizing and classifying different types of physical behavior, such as dynamic and static activities, including walking and running.

2.1.7 Temperature (TEMP)

The Infrared Thermopile sensor is responsible for capturing peripheral skin temperature, which reflects the thermal state at the skin's outer surface rather than the body’s core temperature [20]. This measurement is generally more sensitive to external influences and shows greater fluctuations. Typical values range from 25°C to 35°C and are affected by variables such as surrounding temperature, the efficiency of local blood circulation, and the person's current physical exertion level.

2.2 Machine Learning Algorithms

Machine learning techniques are typically divided into three primary types: **supervised learning, unsupervised learning, and reinforcement learning** [21]. In the supervised paradigm, each input is matched with a known output label, allowing the model to learn predictive relationships and apply them to unfamiliar data. This method is often described as learning with explicit instruction or guidance. In contrast, unsupervised learning is applied to datasets without labeled outcomes, aiming to discover intrinsic data patterns such as groupings or correlations. Reinforcement learning differs by involving an agent interacting with an environment, receiving evaluative feedback in the form of rewards or penalties based on its performance, and gradually refining its decision-making strategy. In the context of this thesis, all algorithms explored fall within the supervised learning framework, as a psychologist has manually assigned labels to every data entry. The study emphasizes decision tree-based models due to their proven reliability in similar classification tasks reported in prior literature [16],[22]. Decision trees offer a transparent and structured approach to classification, where nodes reflect decision criteria based on features, branches depict decision outcomes, and leaves represent final predicted outputs. This section outlines the specific algorithms applied to perform classification tasks in the study.

2.2.1 Adaptive Boosting Decision Algorithm

The AdaBoost algorithm [23], also referred to as Adaptive Boosting Decision Tree , is an ensemble learning technique designed to enhance the performance of weak classifiers by combining them into a single, robust model. The process begins with a given dataset in which each instance is initially assigned an equal weight of 1/N, where N represents the total number of training samples.

A weak classifier is trained on this weighted dataset, and its classification error is subsequently computed. Based on this error rate, a corresponding weight is assigned to the classifier, reflecting its contribution to the final model. The instance weights are then updated: weights of misclassified samples are increased, thereby directing greater attention to them in subsequent iterations. Following this, the instance weights are normalized to ensure they sum to one.

This iterative process continues for a predefined number of rounds or until a specified stopping criterion is satisfied. In the final stage, all trained weak classifiers are aggregated using a weighted majority voting scheme, where each classifier's influence is proportional to its accuracy. The resulting strong classifier is then evaluated on a separate test set to assess its predictive performance.

2.2.2 Gradient Boosting Decision Tree

Gradient Boosting Decision Tree (GBDT) [24],[25] is an ensemble learning method that incrementally builds a strong predictive model by sequentially training multiple shallow decision trees. At each iteration, the algorithm applies gradient descent to minimize a specified loss function by fitting new trees to the residual errors of the previously constructed ensemble. This iterative refinement process allows the model to progressively correct its mistakes and improve performance. The final output is a weighted sum of all the individual trees, resulting in significantly enhanced accuracy compared to any single decision tree.

2.2.3 Bagging Decision Tree

The Bagging Decision Tree algorithm [26] belongs to the family of ensemble methods and works by generating several decision trees, each trained on a randomly resampled (bootstrapped) portion of the original dataset. Once trained, the individual models contribute to the final output through a voting or averaging mechanism, which enhances predictive performance. This approach helps to minimize model variance and reduces the risk of overfitting by leveraging the diversity of the tree ensemble.

2.2.4 Random Forest

Random Forest [27] is a robust ensemble technique that enhances predictive accuracy by generating a collection of decision trees, each trained independently. It extends the standard Bagging Decision Tree approach by introducing stochastic elements not only in data selection, through bootstrapped sampling, but also during the node-splitting process, where a random subset of features is evaluated for optimal splits. This intentional randomness fosters greater model diversity, helping to reduce overfitting and variance, ultimately yielding a more stable and generalizable predictive model.

2.2.5 Extra Trees

Extra Trees [28] is an ensemble learning technique that constructs multiple decision trees to produce a more accurate and stable model. It enhances the randomness introduced in the Random Forest algorithm by adding further variation during tree construction. Each tree is trained on a bootstrapped subset of the data and considers a random subset of features at each node. Unlike Random Forest, however, Extra Trees selects split points at random rather than searching for the optimal ones. This increased randomness helps to lower the risk of overfitting and promotes greater diversity among the trees, ultimately reducing variance and improving the ensemble's overall performance.

2.3 Model Evaluation

To determine the most effective classification algorithm for the purpose of this thesis, it is essential to select a suitable evaluation methodology and relevant performance metrics for comparing the candidate algorithms.

2.3.1 Evaluation Methodology

The evaluation methodology applied to compare the performance of the Machine Learning algorithms is Stratified *k*-fold cross-validation [29], a technique commonly employed in previous studies[4],[5],[6],[7] and [8]. This method ensures that each fold maintains the same class distribution as the original dataset, allowing for balanced representation across all folds. The dataset is split into *k* equal-sized subsets, each preserving the proportion of class labels. Over *k* iterations, each subset is used once as a validation set while the remaining *k-1* subsets form the training set. In each round, the model is trained on the training folds and evaluated on the validation fold, generating performance metrics such as accuracy, precision, and recall. The final performance is derived by averaging these metrics across all *k* folds, providing a thorough and reliable assessment of the model's effectiveness.

2.3.2 Performance Metrics

To calculate the performance metrics [30], four key components are required: true positives, false positives, true negatives, and false negatives, defined as follows:

- 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, within the emotional imagery experiment, specifically in the first two scenarios discussed in Section 1.3, a positive classification refers to patients categorized as having **high levels of stress**, whereas a negative classification indicates patients experiencing **low levels of stress**. In contrast, for the last two scenarios, which are discussed in Section 1.3, a positive classification corresponds to patients with **low levels of psychological flexibility**, while a negative classification refers to those with **high psychological flexibility**.

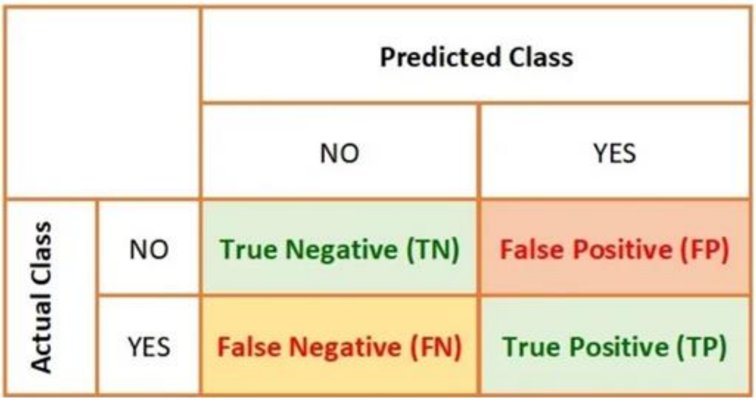
For the EMA experiment, by positive it is meant that the participant is in the category of avoidance or dysfunctional, while negative means that the participant is considered as acceptance or functional. Additionally, false negatives are more vital, in the context of this thesis, than false positives.

For the emotional imagery experiment, minimizing false negatives is of greater importance than minimizing false positives, particularly due to the sensitive nature of the domain, which involves the early detection of psychological distress using physiological features. A false negative, where a subject experiencing high levels of stress or low psychological flexibility is incorrectly classified as not being at risk, may result in the individual not receiving necessary psychological support or intervention. This poses ethical concerns, as undetected distress can escalate and potentially lead to more severe mental health outcomes. On the other hand, false positives, while not ideal, results in additional but generally non-harmful follow-up actions, such as further screening or monitoring.

In the Empatica E4 experiment, the data pertains to healthcare and diagnostic applications. Therefore, it is significantly more critical to avoid classifying a dysfunctional individual as functional, as this could lead to delays in diagnosis and access to necessary treatment. Conversely, if a functional individual is mistakenly classified as dysfunctional, they would likely undergo additional examinations before any treatment or medication is administered, during which a correct diagnosis could still be made. As a result, the consequences of a false negative are more severe than those of a false positive in this context.

Confusion Matrix

The confusion matrix is a square table that presents the four classification outcomes described earlier. Its structure is illustrated in Figure 2.3.



		Predicted Class	
		NO	YES
Actual Class	NO	True Negative (TN)	False Positive (FP)
	YES	False Negative (FN)	True Positive (TP)

Figure 2.3: Binary Confusion Matrix format

Accuracy

Ratio of correct predictions to total predictions.

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

Sensitivity/Recall

Ratio of true positives to total (actual) positives in the data.

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

Precision (PPV)

Ratio of true positives to total predicted positives.

$$\mathrm{PPV} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

Specificity

Ratio of true negatives to total negatives in the data.

$$\mathrm{Specificity} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

F1-Score

Considers both precision and recall. It is the harmonic mean of the precision and recall.

$$\mathrm{F1\ Score} = \frac{2 * (\mathrm{Recall} \ * \ \mathrm{Precision})}{\mathrm{Recall} \ + \ \mathrm{Precision}}$$

NPV

Ratio of true negatives to total predicted negatives.

$$\mathrm{NPV} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FN}}$$

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.

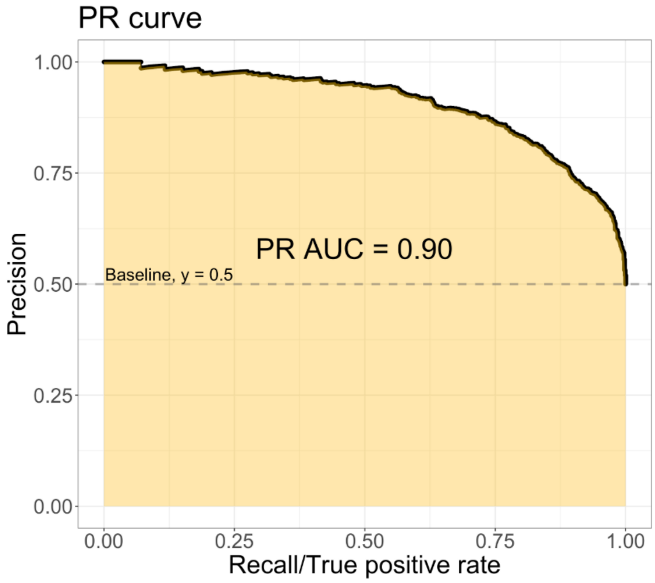


Figure 2.5: Precision-Recall Curve and AUC example

2.4 Monitoring Devices

In this thesis, two wearable biosensing devices are used to collect physiological data: the **Shimmer sensor platform** and the **Empatica E4 wristband**. Both devices are capable of recording signals related to stress and emotional states, such as heart rate and electrodermal activity. The Shimmer device was used for the first experiment, which was an emotional imagery experiment, while the Empatica E4 was used for the Ecological Momentary Assessment (EMA) experiment, each selected based on the specific requirements of the experimental design.

2.4.1 Monitoring Device - Shimmer Sensors

For the emotional imagery experiment presented in Section 3, the monitoring device that was used was Shimmer sensors as illustrated in Figure 2.6. Shimmer sensors constitute a versatile and wearable biosensor platform, specifically designed for monitoring physiological signals in real-time, making them highly suitable for ambulatory biomedical applications [31]. These lightweight (~20 grams), wireless devices provide significant advantages in usability due to their compact form and ease of integration into everyday activities without significant interference or discomfort [31].

Technically, Shimmer sensors feature both Bluetooth and wireless capabilities, enabling efficient data streaming and remote monitoring. They also support SD card logging for scenarios requiring extended data collection periods without continuous streaming. Shimmer sensors include various interchangeable modules (daughter-boards) capable of capturing multiple physiological signals, including Photoplethysmography (PPG), Electrodermal Activity (EDA), Electrocardiography (ECG), and Electromyography (EMG), among others [31]. Specifically, the PPG sensor can measure critical parameters such as Blood Volume Pulse (BVP) and Heart Rate Variability (HRV), while the EDA sensor captures skin conductance levels (SCL), which are indicative of emotional and stress responses [31].

The accuracy and reliability of the Shimmer platform have been extensively validated against well-established laboratory devices, showing comparable results in physiological measurements like ECG, EMG, and GSR (Galvanic Skin Response) [31],[32]. Due to these characteristics, Shimmer sensors have been widely utilized in numerous research domains including stress detection, emotional state assessment, sleep studies, gait analysis, and ambulatory health monitoring [31],[32].

Considering the objectives of this thesis, Shimmer sensors were selected due to their proven effectiveness in machine learning-based classifications of psychological states such as stress and psychological flexibility. Specifically, their capacity to reliably capture physiological signals such as EDA and PPG allows for accurate detection and prediction of patients' stress levels and psychological flexibility in controlled laboratory experiments, making them particularly suitable for this research involving real cancer patients [31].

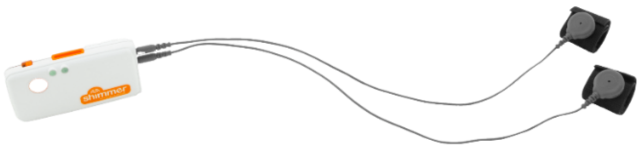


Figure 2.6: Shimmer sensors portable device

2.4.2 Monitoring Device - Empatica E4

For the EMA experiment presented in Section 3, the monitoring device that was used was the Empatica E4 wristband as illustrated in Figure 2.7. The Empatica E4 is a wearable wristband designed specifically for continuous, real-time monitoring of physiological signals in everyday life. Its user-friendly design and compact size make it suitable for long-term wear without discomfort, making it ideal for ambulatory studies and real-world data collection scenarios [33].

This device is primarily developed to facilitate stress detection, emotional monitoring, and health-related studies by accurately capturing physiological signals that are indicative of autonomic nervous system activity and emotional states [34]. It weighs approximately 25 grams, features real-time Bluetooth streaming capabilities, and supports data storage in onboard memory, allowing continuous data collection even when wireless streaming is interrupted.

The Empatica E4 is equipped with several built-in sensors that enable the measurement of a wide range of physiological signals. It includes a photoplethysmography (PPG) sensor, which can capture Blood Volume Pulse (BVP), heart rate (HR), and heart rate variability (HRV), all of which are critical indicators of emotional and physiological states [34][35]. It also features an electrodermal activity (EDA) sensor for monitoring skin conductance, which reflects sweat gland activity and serves as a direct indicator of emotional arousal and stress [34][35]. Additionally, the device includes a three-axis accelerometer for tracking physical movement and activity levels, as well as a skin temperature sensor that records peripheral temperature fluctuations associated with stress or emotional changes.

The Empatica E4 has been extensively validated in a variety of research contexts, including stress detection, emotional state assessment, and health monitoring. It has been particularly effective in psychological research focused on classifying coping behaviors such as acceptance versus avoidance [35]. For the current thesis, which aims to classify real cancer patients into acceptance or avoidance categories based on psychological flexibility metrics, the Empatica E4 is an ideal choice. Its accuracy in capturing nuanced physiological responses and its compatibility with real-time data analysis make it a powerful tool for developing machine learning models capable of supporting clinical decision-making and behavioral interventions.

Figure 2.7: Empatica E4 wristband

Chapter 3 - Data Collection and Previous Work

3.1. Physiological Experiments

The Department of Psychology at the University of Cyprus conducted four distinct experiments: Diagnosis of Experiential Avoidance in Smokers, Diagnosis of Eating Disorders, Diagnosis of Experiential Avoidance for Anxiety, and Functional versus Dysfunctional Coping with Acute Pain. These experiments were conducted at the ACT Health lab and involved volunteer participants. A key difference between these previous studies and the current research is that the present experiments involve real-life cancer patients.

All experiments were connected to Acceptance and Commitment Therapy (ACT), categorizing participants into acceptance (functional) or avoidance (dysfunctional) groups based on their reactions. Acceptance-based strategies encourage individuals to embrace their thoughts and sensations, while avoidance-based strategies involve efforts to avoid, control, or alter uncomfortable thoughts and sensations [12]. Moreover, an individual's classification can vary depending on environmental factors and situational context. For the first three experiments, participants were consistently classified into a single group throughout the procedures. However, this assumption was not applicable in the fourth experiment.

3.1.1 Previous Physiological Experiments

The first experiment targeted experiential avoidance in smokers, aiming to differentiate between smokers employing acceptance versus avoidance coping strategies. Recorded signals included Electrocardiogram (ECG), facial Electromyography (fEMG), and Galvanic Skin Response (GSR), chosen to monitor emotional and stress responses while participants engaged in cognitive tests and watched emotionally neutral or negative videos.

The second experiment addressed emotional regulation related to eating disorders, categorizing participants as low or high risk based on responses to emotionally neutral and disorder-related videos. ECG, fEMG, and GSR signals were captured, alongside self-reported assessments using the Body Image Acceptance and Action Questionnaire (BI-AAQ).

In the third experiment, researchers focused on anxiety, comparing emotional regulation between acceptance and avoidance groups. Participants viewed emotionally provocative images while ECG, GSR, and fEMG (COR, ORB, ZYG muscles) data were collected.

The fourth experiment explored functional versus dysfunctional coping mechanisms for acute pain. Participants underwent pain induction through the Cold Pressor Task (CPT), with coping strategies directed by varied instructions. Behavioral measures (pain tolerance and threshold), psychophysiological data (ECG, EDA, and fEMG), and self-reported psychological assessments were collected. This experiment used both stationary and wearable devices to compare data reliability.

3.1.2 Emotional Imagery Experiment

This is one of the recent experiments, and the one of the two that this thesis focuses on. The experiment involving Shimmer sensors was conducted in a controlled laboratory setting and aimed to measure physiological responses to emotional stimuli before and after a psychological intervention. Each participant completed the experimental protocol twice, once prior to the intervention and once following it. At the beginning of each session, participants were fitted with Shimmer biosensors: two sensors were placed on the index and middle fingers to capture skin conductance, and another sensor was attached to the ear to monitor heart rate through photoplethysmography (PPG). These sensors continuously recorded physiological signals throughout the duration of the experiment.

The procedure began with a five-minute heart rate variability (HRV) assessment during which participants sat quietly and relaxed while baseline heart rate was measured. Following this, participants completed six guided emotional imagery trials. The trials were categorized into two neutral scripts (N1 and N2), two pleasant scripts (J1 and J2), and two unpleasant scripts (F1 and C1). At the start of each trial, participants received a printed script to read and memorize. Once the trial began, participants underwent a 90-second process divided into three 30-second phases: a baseline phase where they counted silently, an imagery phase where they vividly imagined the scenario from the script, and a recovery phase where they relaxed. Event markers were manually inserted by the research assistant during each trial to indicate the start of the HRV period and each imagery trial.

At the conclusion of each trial, participants completed subjective ratings about their experience through some self-report psychological questionnaires before proceeding to the next script . This structured approach allowed researchers to gather psychophysiological data aligned with specific emotional responses under standardized conditions.

The first questionnaire that was considered was the stress subscale of the Depression Anxiety Stress Scale (DASS-S), which measures the individual's perceived level of stress. Higher scores on this subscale indicate greater levels of stress. The second questionnaire was the Acceptance and Action Questionnaire-II (AAQ-II), which measures psychological flexibility. This total score reflects a participant's ability to accept difficult thoughts and emotions while continuing to act according to their values, with lower scores indicating greater flexibility and higher scores suggesting experiential avoidance.

Together, the physiological signals recorded by the Shimmer sensors and the self-report questionnaire data were used to investigate the relationship between emotional processing and psychological flexibility, providing a comprehensive overview of the participants' stress responses and coping mechanisms.

3.1.3 EMA Experiment

This experiment is one of the most recent and is one of the two main experiments discussed in this thesis. Participants entered the Ecological Momentary Assessment (EMA) phase [36]. During this phase, they were provided with wearable psychophysiological monitors and instructed to wear them over the course of three consecutive days. Participants were guided on how to properly use the Empatica E4 wristband, including instructions for wearing, charging, and ensuring continuous data collection.

An app developed by the researchers was installed directly onto each participant's personal smartphone. Through this app, participants were prompted to answer a series of questions multiple times per day. Rather than a fixed schedule, participants were allowed to choose one of three available time ranges for receiving prompts: 8am to 8pm, 9am to 9 pm, or 10 am to 10 pm. Within their chosen range, they received notifications three times a day to complete the questionnaires.

The questions assessed factors such as social context, experiences of physical or emotional pain or stress, and the use of coping strategies. Participants wore the Empatica E4 wristband continuously throughout the day and removed it only at bedtime, when they were instructed to charge the device. Reminder messages were automatically sent every 30 minutes in cases where participants did not initially respond to the prompts. After the three-day period, participants returned the devices to the researchers.

3.2 Related Work

Five prior works examined the data of the experiments regarding smoking, eating disorders, anxiety as well as pain and emotions management. These previous projects were conducted between 2017 and 2023, all focusing on analyzing psychophysiological data collected during experiments involving coping strategies like acceptance and avoidance.

The diploma project by Ch. Galazis in 2017 [4] focused on classifying participants from experiments related to smoking and eating disorders using Random Forest and various machine learning algorithms. The main contribution of this work was identifying the optimal combination of signal features, primarily mean values from each timeframe, for classification. Galazis tested multiple models, such as SVM, Neural Networks, and Adaptive Boosting Decision Tree, with results evaluated across ten different data splits.

The 2018 master thesis by A. Trigeorgi [37] advanced this approach by incorporating ECG time-domain features into the feature set. The study emphasized feature extraction, using Stratified 5-fold cross-validation to evaluate model performance. Similar algorithms were used, and average performance across five splits helped identify the best-performing model.

In 2019, G. Demosthenous built on this foundation by extracting a broader set of ECG-based features and ranking them using Gradient Boosting Decision Tree (GBDT) [5]. The study introduced two data augmentation techniques, Moving Window Methodology and Rectangular Window Methodology, to increase the sample size. It also evaluated five tree-based algorithms and ran each one 100 times per dataset split, enhancing the statistical reliability of results.

The 2022 diploma project E. Georgiou [6] focused on pain-coping strategies in an experiment using the Cold Pressor Task and three devices: BIOPAC, Microsoft Band 2, and Moodmetric Smart Ring. The study emphasized the comparison of physiological data from wearable and stationary devices, the generation of artificial samples via rectangular windows, and the selection of relevant features using Wrapper, Embedded, and Filter Methods.

Finally, the diploma thesis by S. Zeniou in 2023 [8] extended this body of work by applying machine learning to real-time data from the Empatica E4 wristband. This study involved participants from the University of Cyprus who wore the device and responded to questions on an app installed on their smartphones. Physiological signals such as PPG, EDA, temperature, and accelerometer data were analyzed and features were selected using two different methods. Among the tested models, Adaptive Boosting Decision Tree, GBDT, Random Forest, and Extra Trees, the GBDT performed best, achieving 70% accuracy in classifying participants' pain coping strategies. The study also highlighted the importance of HRV features and demonstrated that data collected from wearable devices can yield results comparable to those of stationary systems.

Chapter 4 - Signal Analysis

4.1 Data Selection

The contents of each dataset, the number of patients who participated in each experiment, the method used for categorizing the patients, and the specific data utilized for analysis are among the key details that will be presented in the current section.

4.2.1 Data Selection for Emotional Imagery Phase

A comprehensive data collection process was conducted for the purposes of machine learning. The study (referred to in Section 3.1.2) involved a total of 72 patients with valid data, each of whom was associated with one .csv file. These .csv files contained various physiological measurements, each recorded in a separate column. For instance, the column Shimmer__GSR_Skin_Conductance_CAL corresponded to Electrodermal Activity (EDA), Shimmer__PPG_A13_CAL represented Blood Volume Pulse (BVP), and Shimmer__PPG_IBI_CAL represented the Interbeat Interval (IBI). Each file also included a column for the timestamp of each measurement as well as a column indicating the specific trial during which the measurement was taken.

Additionally, the scores obtained from the questionnaires completed by the patients were collected into a single CSV file. Although multiple scores were gathered, this thesis focuses only on two of them: the DASS score and the AAQ-II score, which were explained in detail in Section 3.1.2. Each score was stored in a separate column, and an additional column contained patient IDs, allowing for clear identification of which patient each score belonged to.

A total of 80 patients participated in the practical component of the experiment, meaning that corresponding physiological data (CSV files) were available for them. However, 8 of these participants did not complete the required questionnaires during the experimental procedure. Additionally, there were 5 patients who completed the questionnaires but did not participate in the practical component. Therefore, from the 85 patients who were involved in the study, either by participating in the practical session, completing the questionnaires, or both, a total of 72 participants were retained after the data cleaning process described in Section 5.4.1. These 72 patients had both physiological data and completed questionnaire responses, making them eligible for further analysis.

Four different scenarios were examined, based on two different thresholds for each score, to classify patients into categories of low/high stress and low/high psychological flexibility as described in section 1.3.

In the first scenario, where the threshold was 7, patients who scored 7 or below were classified as having a "low level of stress," while those who scored above 7 up to the maximum possible score of 21 were classified as having a "high level of stress." In the second scenario, using the median value of 5 as the threshold, patients who scored 5 or below were placed in the "low level of stress" category, while those who scored above 5 up to 21 were classified as having a "high level of stress".

For the AAQ-II score, classification was also performed using both a literature-based threshold (24) and the median value of the data (13.5). In the first scenario, where the threshold was 24, patients who scored above 24 were classified as having a "low level of psychological flexibility," while those who scored 24 or below were placed in the "high level of psychological flexibility" category. In the second scenario, using the median threshold of 13.5, patients who scored above 13.5 were categorized as having a "low level of psychological flexibility," and those who scored 13.5 or below were considered to have a "high level of psychological flexibility."

The primary data points extracted for analysis were the values recorded during each trial in which patients participated. This approach allowed for a focused investigation of the relationship between physiological measurements and the patients' pain coping strategies.

4.1.2 Data Selection for EMA Phase

A comprehensive data collection process was carried out to support the machine learning analysis. As outlined in Section 3.1.3, the study involved 26 patients, each of whom was associated with six data files containing various physiological metrics: Acceleration (ACC), Interbeat Interval (IBI), Heart Rate (HR), Temperature (TEMP), Electrodermal Activity (EDA), and Blood Volume Pulse (BVP). These data were collected over a span of three consecutive days. Alongside the physiological recordings, participants completed a short questionnaire three times per day, and the exact time of each response was recorded. The primary question included in the questionnaire was: “If need be, I can let unpleasant thoughts and experiences happen without having to get rid of them immediately?”

Participants could respond to this question with a score ranging from 1 to 5. Since there was no proposed threshold value in the existing literature for this specific experimental setup, the categorization of responses was based on the median value of participants’ answers to this question. The median score was 4, therefore, responses of 4 or 5 were categorized as “acceptance”, while scores of 3 or lower were categorized as “avoidance”.

The first step in the data processing involved identifying the exact time each participant answered the coping-related question. Following this, the corresponding physiological data files were accessed, and the relevant timestamp was located. From these files, the physiological values recorded during the five-minute window preceding the participant’s response were extracted. This method enabled a targeted investigation into how physiological activity correlates with different coping strategies in the context of pain management.

4.2 Feature Extraction

Raw psychophysiological signals are not suitable for directly training machine learning algorithms. Therefore, it is essential to extract meaningful features from these signals. This section outlines the specific features derived from each type of signal.

4.2.1 Features Extracted from PPG Signal

Time-domain measures are fundamental metrics in heart rate variability (HRV) analysis. They quantify the variability in time intervals between consecutive heartbeats, known as NN (normal-to-normal) intervals. To understand what NN intervals are, it is necessary to revisit the concept of RR intervals from section 2.1.1. An RR interval refers to the time between two successive R-peaks in the ECG signal. Consequently, an NN interval represents the time between successive **normal** R-peaks, specifically, R-peaks that are not affected by artifacts [38]. To extract features from the raw PPG signals, the Python libraries Flirt [39] and either HeartPy [49] or NeuroKit2 [50], depending on the dataset associated with each experiment, were employed. These libraries enabled efficient preprocessing and feature computation tailored to the characteristics of the collected physiological data. The process begins with the computation of RR intervals, as outlined in Table 4.1, along with the calculation of the differences between successive RR intervals (RRdiff) and their squared differences (RRsqdiff).

Short name	Clarification	Equation
RR	It is the time interval between consecutive R-peaks, measured in milliseconds.	RR = (diff(R_{peaks})/sf)*1000, where sf is the sampling frequency
RRdiff	The absolute value of the differences between successive RR intervals.	$RR_{diff} = \|diff(RR)\|$
RRsqdiff	The squared differences between consecutive RR intervals.	$RR_{sqdiff}=RR^2_{diff}$

Table 4.1: Metrics used to represent time-domain features

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

Short name	Clarification	Equation	Unit
IBI	Inter-Beat Intervals: The average of RR intervals, representing the time between consecutive R-waves, measured in milliseconds.	$IBI = RR$	ms
BPM	Beats Per Minute: The average number of heart beats per minute.	$\text{BPM} = \frac{60000}{RR}$	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
RMSSD	Root Mean Square of Consecutive RR Interval Differences	$RMSSD = \sqrt{\frac{1}{N-1}\sum(RR_{diff_i})^2}$	ms
pNN20	The proportion of consecutive NN interval differences greater than 20 ms relative to the total number of consecutive NN intervals.	$pNN20 = \frac{count(diff(RR) > 20ms)}{count(diff(RR))}$, where count(X) gives the number of elements in X	%
pNN50	The proportion of consecutive NN interval differences exceeding 50 ms compared to the total number of 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 (MAD) 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: Metrics used to represent time-domain features

Additionally, several frequency-domain features were extracted, as presented in Table 4.3. These features are based on the power spectral density (PSD) of the RR intervals, which is computed using techniques such as the Fast Fourier Transform (FFT) or autoregressive modeling. The total power derived from the PSD is then segmented into distinct frequency bands[38].

Short name	Clarification
VLF(Very Low Frequency)	Power within the very low frequency (VLF) band, typically ranging from 0.0033 to 0.04 Hz.
LF (Low Frequency)	Power within the low frequency (LF) band, typically ranging from 0.04 to 0.15 Hz.
HF(High Frequency)	Power within the high frequency (HF) band, typically ranging from 0.15 to 0.4 Hz.
LF/HF Ratio	The ratio of LF power to HF power.
LFnU (LF normalized units)	LF power expressed in normalized units, calculated by dividing the LF power by the total power minus the VLF power, and then multiplying the result by 100.
HFnU(HF normalized units)	HF power expressed in normalized units, calculated by dividing the HF power by the total power minus the VLF power, and then multiplying the result by 100.

Table 4.3: Metrics used to represent frequency domain features

Moreover, non-linear features are extracted from the Poincaré plot, a scatterplot that maps each RR interval against the subsequent one [40], as illustrated in Table 4.4.

Short name	Clarification
SDI	The standard deviation of the points perpendicular to the line of identity in the Poincaré plot, representing the short-term variability of the heart rate.
SD2	The standard deviation of the points along the line of identity in the Poincaré plot, indicating the long-term variability of heart rate.
SD2/SD1 Ratio	The ratio of SD2 to SD1.

Table 4.4: Metrics used to represent non-linear features

4.2.2 Features Extracted from EDA Signal

Electrodermal Activity (EDA) signals allow for the extraction of four primary types of features: time-domain features, frequency-domain features, time-frequency domain features, and Mel-frequency cepstral coefficients (MFCCs) [41]. In this thesis, the focus is placed on statistical features, given their extensive use and validation in existing literature.

The EDA signal is generally decomposed into two distinct components. The first is the Tonic component, also known as Skin Conductance Level (SCL), which reflects slow-varying, baseline changes in skin conductance. This component is typically associated with an individual's overall level of arousal or stress over extended periods. The second is the Phasic component, referred to as Skin Conductance Response (SCR), which captures rapid and transient fluctuations in skin conductance triggered by specific stimuli or events. This component represents short-term or immediate physiological responses.

For feature extraction, as presented in Table 4.5, the Flirt [39] Python library was employed. This tool facilitated the systematic derivation of relevant statistical features from the EDA signal for further analysis.

Short name	Clarification
tonic_mean	Average value of SCL.
phasic_mean	Average value of SCR.

Table 4.5: Extracted statistical features from the EDA signal

4.2.3 Features Extracted from 3-axis Accelerometer

Finally, the features extracted from the three-axis accelerometer data are presented in Table 4.6. In the context of human activity monitoring, this data is utilized to capture body movement patterns[42]. As with previous signals, the Flirt [39] Python library was employed for feature extraction.

Short name	Clarification
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

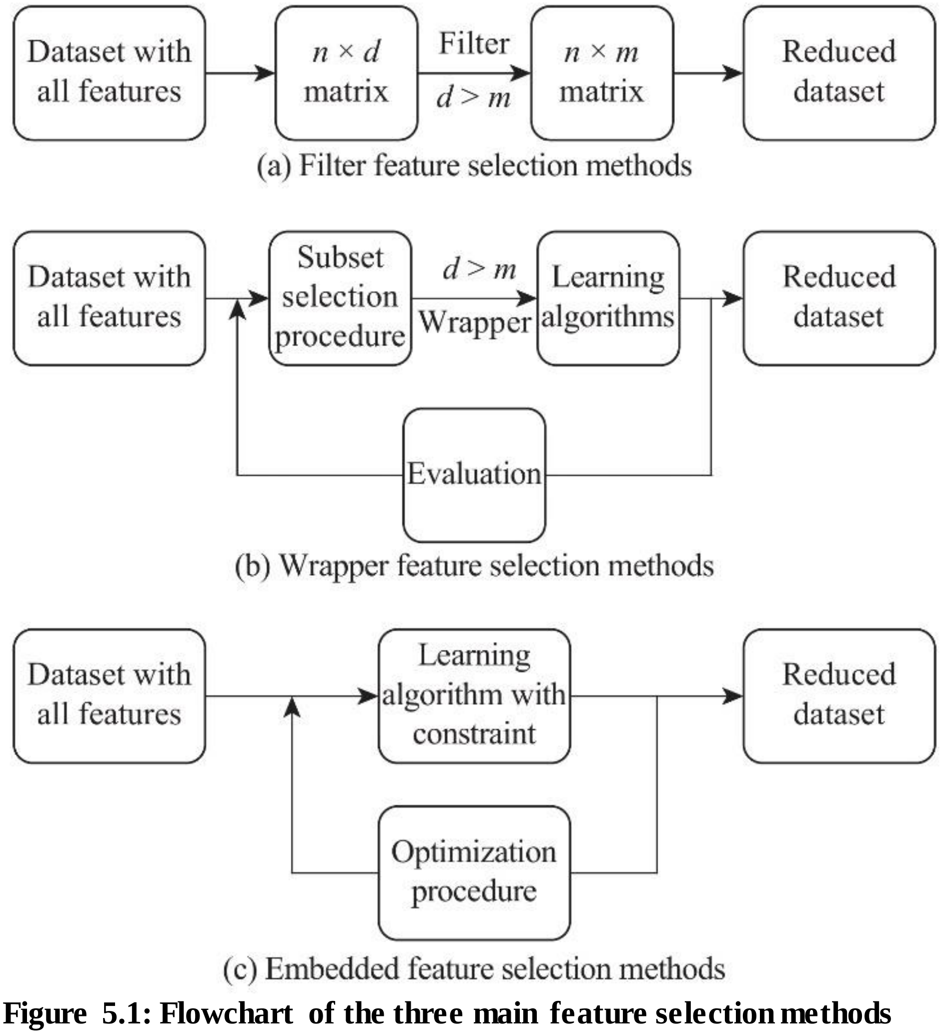
Chapter 5 - Feature Selection

5.1 Feature Selection Techniques

Feature selection is a fundamental step in the machine learning pipeline that involves selecting a subset of relevant features from the original features dataset to improve the performance of predictive models. By eliminating irrelevant, redundant, or noisy features, feature selection helps to enhance model generalization and reduce the risk of overfitting, leading to more accurate predictions [43]. Additionally, by decreasing the number of input variables, the training time and computational complexity of machine learning algorithms are significantly reduced, making the modeling process more efficient [43].

Beyond computational efficiency, feature selection improves model interpretability, particularly in domains such as healthcare, where understanding the influence of individual features on the prediction outcome is essential [43]. Furthermore, it reduces the storage requirements and enhances data quality by focusing only on the most informative variables, which may offer deeper insights into the underlying problem being analyzed [43]. For the purposes of this thesis, in physiological signal analysis, feature selection helps identify the most relevant biomarkers contributing to stress or emotional state detection, thus streamlining the modeling process and improving the robustness of results. Recognizing the critical role of this step in enhancing predictive accuracy, a structured approach to feature selection was adopted.

Specifically, and within the scope of supervised learning, various feature selection techniques have been employed, following the methodology applied by Sotiris Zeniou in his 2023 thesis project [8]. The feature selection process was primarily based on three main methodological categories: Wrapper methods, Filter methods, and Embedded methods. These categories will be analyzed and compared in the following sections, with the aim of evaluating their effectiveness. Figure 5.1 illustrates the core concept behind each category, which will be discussed in more detail throughout this chapter.



(a) Filter feature selection methods

(b) Wrapper feature selection methods

(c) Embedded feature selection methods

Figure 5.1: Flowchart of the three main feature selection methods

5.1.1 VIF Technique

The first technique applied before implementing the three feature selection methods was the Variance Inflation Factor (VIF) [44], which is used to detect multicollinearity among the features. Multicollinearity refers to a condition where two or more predictor variables are highly correlated, potentially leading to unstable model coefficient estimates, reduced interpretability, and unreliable feature importance scores. The VIF metric quantifies the extent to which the variance of a regression coefficient is inflated due to multicollinearity with other predictors in the model. A VIF value of 1 indicates no correlation, while significantly higher values reflect increasing levels of multicollinearity. Although thresholds may differ across applications, a commonly accepted rule of thumb is that VIF values exceeding 10 indicate problematic multicollinearity, and such features should be considered for removal. This threshold was adopted in the thesis project of Sotiris Zenios [8] and was likewise employed in the current study. By applying this approach, only the most relevant and informative features were retained, ultimately contributing to a more reliable and robust predictive model.

5.1.2 Wrapper Methods

Wrapper methods for feature selection represent a category of machine learning techniques aimed at identifying the most relevant subset of features for a given predictive model. These methods evaluate the contribution of feature combinations by iteratively building and assessing models, ultimately selecting the subset that achieves optimal performance based on a predefined evaluation metric [45]. Common strategies employed in wrapper methods include forward selection, backward elimination, and recursive feature elimination (RFE).

In this thesis, following the methodology applied in the thesis project of Sotiris Zenios [8], the Recursive Feature Elimination (RFE) method was utilized in combination with a RandomForestClassifier from the Scikit-learn library. The dataset was initially preprocessed and the target labels were separated from the predictor variables. RFE was then applied by recursively training the Random Forest model and eliminating the least important feature at each iteration, based on the internal importance scores of the estimator. This process continued until the optimal number of features was reached. The selected subset of features was extracted using the *fit.get_support()* function and used in the final model evaluation.

This approach enabled the identification of a highly informative subset of features tailored to the classifier, thus improving model accuracy and stability. RFE, as a wrapper method, provided a systematic and performance-driven strategy for refining the feature set, making it particularly suitable for the high-dimensional and complex nature of physiological signal data analyzed in this thesis.

5.1.3 Filter Methods

In recent years, filter methods have become increasingly popular in the field of feature selection for machine learning, largely due to their simplicity and computational efficiency. These methods assess the intrinsic properties of the dataset, such as correlation and mutual information, to evaluate the relevance of each feature independently of any specific learning algorithm. By discarding irrelevant or redundant features, filter methods help to reduce dataset dimensionality, thereby improving model performance and reducing the risk of overfitting [45]. Additionally, they contribute to enhanced model interpretability and generalization, both of which are essential for producing reliable and robust results.

In this thesis, and in line with the methodology applied in the thesis project of Sotiris Zenios [8], the filter-based method **SelectKBest** was employed to identify the most relevant features from the dataset. The technique ranks features according to their individual scores calculated from univariate statistical tests, independently of the learning algorithm used for classification. Specifically, the *f_classif* scoring function was selected, which is based on the ANOVA F-value. This function evaluates the relationship between each continuous input feature and the categorical target variable by comparing the variance between group means relative to within-group variance.

After computing the F-scores, the top *k* features with the highest values were selected and retained for model training. This approach not only allowed for a significant reduction in feature space but also helped to preserve the most informative attributes, contributing to improved model efficiency and predictive performance. The method's independence from specific classifiers also made it particularly suitable as a general-purpose preselection step for downstream machine learning models.

5.1.4 Embedded Methods

Embedded methods for feature selection incorporate the selection of relevant features directly into the model training process. Unlike filter methods, which evaluate features independently of the learning algorithm, and wrapper methods, which assess feature subsets by repeatedly training models, embedded methods perform selection during model construction. This integrated approach enables the model to inherently consider feature interactions and dependencies, often resulting in improved predictive performance and reduced computational complexity. Common examples of embedded methods include regularization-based techniques like Lasso (L1 penalty), which eliminate irrelevant features by reducing their coefficients to zero, as well as tree-based models such as **Random Forest**, **Gradient Boosting Decision Tree** and **Extra Trees**, which evaluate and rank features based on their contribution to decision splits during training [46].

In the context of this thesis, and following the methodology adopted in the thesis project of Sotiris Zenios [8], embedded methods were implemented using two different ensemble learning models: **GradientBoostingClassifier** and **ExtraTreesClassifier**, both from the Scikit-learn library in Python. The process began with standard data preprocessing, where the dataset was cleaned and the target labels were separated from the predictor variables. The data was then split into training and testing sets to facilitate performance evaluation. For the **Gradient Boosting** approach, the classifier was first trained on the training set. Then, **permutation importance** was applied to measure the relevance of each feature. This technique involves shuffling the values of individual features and observing the corresponding drop in model performance, thus quantifying each feature's predictive contribution. Features that led to the largest performance decline were considered the most informative [47].

For the **Extra Trees** approach, feature importance was extracted directly from the trained model using its built-in feature_importances_attribute. The **ExtraTreesClassifier** is an ensemble method that builds multiple randomized decision trees and averages their predictions. During training, it computes the importance of each feature based on how much it reduces impurity (such as Gini impurity or entropy) across all the decision trees. This method is particularly robust to overfitting and efficient in high-dimensional spaces, making it highly suitable for complex physiological datasets.

By applying these two embedded methods, the most informative features were effectively identified across all scenarios. This dual approach not only strengthened the reliability of the feature selection process but also improved model robustness, interpretability, and generalization, core objectives of this research. The use of both permutation-based and impurity-based importance metrics further reinforced the consistency and validity of the selected features.

5.2 SelectedFeatures for the Initial Datasets

In this section, the features selected through the feature selection process are presented. The selected features for each initial dataset will be presented in separate sections, corresponding to the two experimental phases. For the first phase (emotional imagery experiment phase), features were selected for four distinct datasets, each corresponding to one of the four scenarios described in Section 1.3. For the second phase (EMA experiment phase), feature selection was performed for the single scenario that was conducted.

It is important to note that some features appearing in both experimental phases are equivalent in meaning, although they are labeled differently. As noted in Tables 4.2 and 4.3, BPM refers to hrv_mean_hr, IBI to hrv_mean_nni, LFnU to lf_perc, and HFnU to hf_perc.

5.2.1 SelectedFeatures for the Emotional Imagery Experiment Initial Dataset

The process was applied to the four distinct variations of the original dataset derived from the emotional imagery experiment, each corresponding to a different classification scenario as described in section 1.3.

To identify the most informative features for each dataset variation, a combination of feature selection techniques was employed, including **SelectKBest, RandomForestClassifier**, **Gradient Boosting Decision Tree**, and **Extra Trees**. The results of these methods are summarized in the following table, which highlights the selected features across the different scenarios, providing insights into how feature relevance varies depending on the classification strategy and threshold used.

	RandomForest Classifier	**SelectKBest**	**Gradient Boosting Decision Tree**	**Extra Trees**
Scenario 1 Dataset	hrv_mean_hr, hrv_mean_nni, hrv_pnni_50, hrv_pnni_20, hrv_hf, tonic_mean, phasic_mean, bpm, sdsd,rmssd, pnn20,pnn50, sd1,sd1/sd2, lf_perc	tonic_mean, sdsd,rmssd, pnn20,sd1, sd1/sd2,lf_nu, hf_nu	hrv_pnni_50, hrv_mean_nni, pnn20	hrv_mean_hr, pnn20, tonic_mean, hrv_mean_nn, hrv_pnni_20, phasic_mean
Scenario 2 Dataset	hrv_mean_hr, hrv_mean_nni, hrv_sdnn, hrv_pnni_50, hrv_pnni_20, hrv_hf, hrv_lf_hf_ratio, tonic_mean, phasic_mean, bpm, rmssd, pnn20,pnn50, sd1, lf_perc	sdnn,sdsd, rmssd, pnn50,hr_mad, sd1, sd2, s	hrv_mean_hr, hrv_sdnn, phasic_mean, hrv_mean_nni, lf_perc, hrv_lf_hf_ratio, sdnn	hrv_mean_nn, hrv_mean_hr, tonic_mean, ibi,bpm, phasic_mean

Scenario 3 Dataset	hrv_mean_hr, hrv_mean_nni, hrv_rmssd, hrv_pnni_20, hrv_lf, hrv_lf_hf_ratio, tonic_mean, phasic_mean, bpm, ibi, sdnn, sdsd,rmssd, lf_perc, hf_perc	tonic_mean, sdnn,sdsd, rmssd, pnn20,pnn50, sd1, sd2	vlf_perc, hrv_hf,bpm, ibi, lf	tonic_mean, bpm, hrv_mean_hr, phasic_mean, pnn50
Scenario 4 Dataset	hrv_mean_hr, hrv_mean_nni, hrv_rmssd, hrv_sdnn,hrv_lf, hrv_hf, tonic_mean, phasic_mean, bpm, ibi, sdsd, rmssd, pnn20,pnn50, sd1/sd2	hrv_rmssd, hrv_sdsd, hrv_pnni_50, sdsd,rmssd, pnn50,hr_ma, sd1	**(No features printed in the output for Gradient Boosting in this case.)**	bpm, hrv_mean_nn, hrv_mean_hr, ibi,sd1/sd2, hr_mad

Table 5.1: SelectedFeatures of emotional imagery phase

Regarding all the algorithms used, the selected features were primarily derived from the HRV and EDA signals. This contrasts with the feature selection observed in the scenarios related to Sotiris Zeniou [8], where, following the data cleaning process, features were also selected from the ACC signal, in addition to HRV and EDA.

This difference can be explained by the nature and setting of the Emotional Imagery Experiment, which was conducted in a controlled laboratory environment using Shimmer sensors. In this context, the selected features were mainly derived from HRV and EDA signals, as the clean and stable conditions allowed for high-quality recordings of these physiological indicators. Participants remained seated and stationary during the guided imagery tasks, which significantly reduced movement-related noise and minimized the relevance of accelerometer (ACC) data. Moreover, the use of structured, emotionally charged scripts triggered targeted psychological responses, such as stress or flexibility, which were more accurately captured through HRV and EDA measures. In contrast, the experiment conducted by Sotiris Zeniou [8] took place in real-life conditions using the Empatica E4 device, where participants were constantly moving and responding to ecological momentary assessments. As a result, ACC features became more informative in that setting, offering valuable insights into physical activity patterns that could be linked to stress or avoidance behaviors. This explains why ACC signals played a more prominent role in feature selection in Sotiris Zenios work [8], whereas HRV and EDA dominated in the lab-based emotional imagery experiment.

5.2.2 SelectedFeatures for the EMA Experiment Initial Dataset

This section presents the features selected from the initial EMA dataset using different feature selection techniques, namely Random Forest Classifier, SelectKBest, Gradient

Boosting Decision Tree, and Extra Trees. The most important features identified by each method are listed below in Table 5.2, highlighting the physiological signals and heart rate variability (HRV) metrics most relevant for classification in the EMA phase.

	RandomForest Classifier	**SelectKBest**	**Gradient Boosting Decision Tree**	**Extra Trees**
Original dataset	value_y_mean, value_z_mean, vlf, num_ibis	value_x_mean, value_y_mean, value_z_mean, vlf, lf/hf,tonic_mean, phasic_mean, num_ibis	(No features printed in the output for Gradient Boosting in this case.)	value_y_mean, value_z_mean, value_x_mean, hrv_mean_nni, sd1/sd2, hrv_sdnn

Table 5.2: SelectedFeatures of EMA phase

Regarding all the algorithms used, it appears that the selected features were primarily derived from the ACC signal. This differs from the feature selection observed in the scenarios related to Sotiris Zeniou [8], where, following the data cleaning process, features were mainly selected from the HRV and EDA signals, in addition to the ACC signal.

This difference can be attributed to the specific conditions and population of the EMA experiment. Conducted in real-world, ambulatory settings using the Empatica E4 device, the EMA experiment involved cancer patients who wore the device during their daily routines. As a result, the HRV and EDA signals were often affected by noise from continuous motion, changes in posture, and varying environmental conditions, making them less reliable and consistent. In contrast, ACC features became more informative in capturing patterns of physical activity and behavioral responses, such as restlessness or avoidance-related movements, which may be linked to psychological states.

Moreover, the EMA data collection was event-based, relying on participants' responses to self-reports triggered throughout the day. These unstructured, real-time windows emphasized the relevance of body movement context, where ACC-derived features, such as *value_x_mean* and *value_z_mean*, proved more stable and discriminative compared to HRV or EDA.

On the other hand, in Sotiris Zeniou's experiment [8], the participants were university student volunteers in a less physically demanding or variable environment. These participants typically follow more structured daily routines (e.g., attending lectures, studying), which are less physically intensive compared to the diverse and often unpredictable routines of cancer patients. This allowed for cleaner physiological recordings and better retention of HRV and EDA signals after preprocessing. Consequently, HRV and EDA features played a more significant role in his feature selection process. In contrast, the daily activities and potential physical and emotional burden faced by real cancer patients in the EMA study likely contributed to greater signal noise and variability, elevating the discriminative value of ACC features in the classification process.

5.3 Optimized SelectedFeatures for the Initial Datasets

To further optimize the results that will be analyzed in chapter 6 and to improve the performance of the predictive models, we combined the four-feature selection methods applied and selected the features that appeared most frequently across them. The table below presents, for each scenario, the common features with the highest number of occurrences identified by the four methods.

5.3.1 Optimized SelectedFeatures for the Emotional Imagery Experiment Initial Dataset

This section summarizes the optimized feature sets selected for each scenario of the emotional imagery experiment after applying feature selection techniques. The selected features, shown in Table 5.3, represent the most relevant physiological signals and heart rate variability (HRV) metrics that contributed to improving the classification performance across the different scenarios.

	Selectedfeatures
Scenario 1 Dataset	hrv_mean_nni, tonic_mean, pnn20
Scenario 2 Dataset	hrv_mean_hr, hrv_mean_nni, phasic_mean
Scenario 3 Dataset	tonic_mean, bpm
Scenario 4 Dataset	hrv_rmssd, sdsd ,rmssd ,pnn50, hr_mad, hrv_mean_hr, hrv_mean_nni, bpm, ibi, sd1/sd2

Table 5.3: Optimized SelectedFeatures of emotional imagery phase

5.3.2 Optimized SelectedFeatures for the EMA Experiment Initial Dataset

Similarly, this section presents the optimized features selected for the EMA experiment dataset. The features listed in Table 5.4 highlight the most influential signals used to enhance the model's ability to distinguish between acceptance and avoidance behaviors.

	Selectedfeatures
Original Dataset	'value_x_mean', 'value_y_mean', 'value_z_mean', 'tonic_mean', 'vlf'

Table 5.4: SelectedFeatures of EMA phase

5.4 Data Cleaning

To create a valid and meaningful dataset for analysis for each phase a data cleaning process was implemented. A main criterion guided the data cleaning process for each phase.

5.4.1 Data Cleaning for Emotional Imagery Phase

Each patient who answered the questionnaire should also have a csv file with measurements from the signals output by the shimmer sensor as described in section
As previously mentioned in Section 4.1.1, each patient had a corresponding csv file containing the signal measurements from the Shimmer sensor for each of the six trials. In this study, the combination of a patient and a specific trial was treated as an individual sample. Therefore, for each patient, six distinct samples were generated. If a patient had not completed any questionnaires, all six potential samples were excluded from the final dataset. Additionally, if a patient lacked signal data for a specific trial, the sample corresponding to that trial was also removed.

5.4.2 Data Cleaning for EMA Phase

Matching the time a patient answered a questionnaire with a time from the Empatica E4 watch measurements:
The data cleaning process implemented in this phase focuses on ensuring that only valid and meaningful samples are retained for analysis. For each unique sample combination of patient and questionnaire, the script attempts to extract physiological signals from the five-minute window preceding the questionnaire timestamp. If no corresponding signal data is found within that time frame for any of the expected modalities (e.g., EDA, BVP, ACC), the respective sample is excluded. Furthermore, if the folder corresponding to a specific patient does not exist, then all potential samples associated with that patient, namely, all possible combinations of patient ID and questionnaire ID, are removed from the dataset. This step ensures that only patient entries with valid and accessible physiological signal data are retained for further analysis, thereby maintaining the integrity and reliabilityof the dataset.

The data cleaning process applied in both experimental phases, emotional imagery and EMA, was essential for ensuring the integrity, consistency, and quality of the dataset. In both cases, strict criteria were implemented to remove incomplete, missing, or irrelevant data entries, such as samples lacking physiological signal measurements or corresponding questionnaire responses. By eliminating such entries, the dataset was refined to include only valid and meaningful samples suitable for supervised learning. This process significantly reduced noise prevented data imbalance caused by invalid samples and ensured that the machine learning models were trained on reliable and contextually relevant information. Overall, the data cleaning phase contributed to improving the robustness, interpretability, and generalizability of the results, making it a critical step in the overall methodology.

Chapter 6 - Classification

6.1 Classification Process

The classification process represents a critical phase of this thesis project, as it directly influences the evaluation of the predictive capabilities of machine learning models in identifying psychological states, such as stress levels and psychological flexibility in cancer patients during guided imagery tasks, and coping strategies (acceptance vs. avoidance) during real-world daily assessments. This step is pivotal in assessing the effectiveness of various machine learning (ML) algorithms and determining which configurations yield the most reliable results across the defined experimental scenarios.

To ensure an unbiased evaluation of the models and avoid overfitting, each dataset was split into training and testing subsets following an 80/20 ratio. This preparation step allows for robust model validation, ensuring that the trained models generalize well to unseen data and that the evaluation metrics truly reflect the models' predictive capabilities.

For the classification task, five supervised machine learning algorithms were utilized: **Gradient Boosting Decision Tree**, **Adaptive Boosting Decision Tree**, **Bagging Decision Tree**, **Random Forest**, and **Extra Trees**. Each of these classifiers was trained on the training dataset using a set of features that had been carefully selected through an optimized feature selection process detailed in Section 5.3. The training phase also involved hyper parameter tuning to enhance the predictive power of each algorithm. These optimizations were performed with the goal of tailoring the models to the underlying structure of the data while preventing overfitting.

The performance of each model was subsequently evaluated on the test set using standard classification metrics, including **accuracy**, **F1-score**, **recall (sensitivity)**, **precision**, **area under the curve (AUC)**, and **specificity**. These metrics provided comprehensive insights into the models' ability to correctly classify patients into distinct psychological categories. In the emotional imagery experiment, this involved distinguishing between **high and low stress levels** based on the DASS score and **high and low psychological flexibility** based on the AAQ-II score. In the EMA experiment, the classification task focused on identifying whether the patient was adopting a **functional (acceptance)** or **dysfunctional (avoidance)** coping response to real-time experiences of stress or pain.

The **primary goal** of this classification process was twofold. First, it aimed to compare the predictive performance across different datasets, which correspond to the four scenarios from the emotional imagery experiment (based on different cut-offs for DASS and AAQII scores) and one scenario from the EMA experiment (based on real-time acceptance vs. avoidance responses). Second, it sought to compare the effectiveness of the five machine learning algorithms in these various settings to determine which method performs best for each case.

To further improve classification outcomes, **oversampling techniques** were applied to address potential class imbalances in the datasets. Specifically, three oversampling strategies were examined: **Random Oversampling**, **SMOTE (Synthetic Minority Oversampling Technique)**, and **ADASYN (Adaptive Synthetic Sampling)**. These techniques will be explained in detail in section 6.3.

Another significant objective of the classification process was to **evaluate the impact of these oversampling techniques** on model performance. By comparing the results before and after oversampling, the study assessed how each technique influenced key evaluation metrics, particularly recall and precision, and whether the inclusion of these techniques led to more balanced and reliable predictions. This analysis was essential in understanding the full potential of machine learning models to accurately identify **maladaptive coping mechanisms in cancer patients**, whether these appear in a controlled setting (as in the emotional imagery experiment) or during daily life (as in the EMA phase).

Overall, this classification framework, aligned with the methodology used by Sotiris Zeniou in his 2023 thesis project [8], provides a systematic and replicable pipeline for analyzing psychophysiological and self-reported data. It demonstrates the utility of machine learning in detecting stress levels, psychological flexibility, and coping styles, offering a foundation for future applications in personalized interventions for cancer care.

6.2 Classifier Selection

As we move into the Classifier Selection chapter, the focus shifts to a detailed comparative analysis of the five distinct scenarios examined throughout this study. A key component of this analysis is understanding the differences in sample distribution across these scenarios, as this factor significantly influences the performance of the machine learning algorithms applied. To support a clear and comprehensive evaluation, Table 6.1 presents the number of patients and Table 6.2 presents the number of samples corresponding to each scenario of each phase (emotional imagery or EMA phase). This overview provides the foundation for the following discussions and assessments of the classifiers, ultimately guiding the selection of the most suitable model based on the unique characteristics of each scenario.

Scenarios	Low level of stress	High level of stress	Low level of psychological flexibility	High level of psychological flexibility
Scenario 1: Stress level (DASS) score, with 7 as the cut off between low and high; the cut off was decided based on prior literature.	**126**	**305**	**-**	**-**
Scenario 2: Stress level (DASS) score, with 5 as the cut off; this cut off was based on the median of our data, provides more balanced categories.	**197**	**234**	**-**	**-**
Scenario 3: Psychological flexibility (AAQII) score, with 24 as the cut off between low and high, based on prior literature.	**-**	**-**	**89**	**342**
Scenario 4: Psychological flexibility (AAQII) score, with 13.5 as the cut off, based on the median.	**-**	**-**	**215**	**216**

Table 6.1: Distribution of Samples Across Scenarios of Emotional Imagery Phase

Scenarios	Avoidance	Acceptance
Empatica E4 experiment: Avoidance vs Acceptance score, with 4 as the cut-off, based on the median	**31**	**76**

Table 6.2: Distribution of Samples in EMA Phase Scenario

Following the data cleaning procedures for each phase, as described in Sections 5.4.1 and 5.4.2 respectively, the final number of valid samples retained was 431 for the Emotional Imagery phase and 107 for the EMA phase. Specifically, 431 out of a possible 510 samples were retained for the Emotional Imagery phase, had all entries been valid, while for the EMA phase, 107 out of 225 potential samples were included in the final analysis.

6.2.1 Scenario 1 of Emotional Imagery Phase Dataset

In Scenario 1, which aimed to classify patients according to their stress levels based on a DASS score threshold of seven (7), the original classification results demonstrated several important performance characteristics across the five evaluated machine learning algorithms as illustrated in Table 6.3.

	Gradient Boosting Decision Tree		Adaptive Boosting Decision Tree		Bagging Decision Tree		Random Forests		Extra Trees	
	Value	SD	Value	SD	Value	SD	Value	SD	Value	SD
Accuracy	0.80	0.05	0.77	0.05	0.80	0.05	0.80	0.05	0.81	0.05
F1-score	0.72	0.07	0.71	0.07	0.75	0.06	0.73	0.07	0.75	0.07
Recall-sensitivity	0.91	0.05	0.88	0.06	0.89	0.06	0.90	0.05	0.93	0.05
Precision(PPV)	0.82	0.04	0.82	0.04	0.84	0.04	0.83	0.04	0.83	0.04
AUC	0.84	0.06	0.80	0.08	0.85	0.06	0.85	0.06	0.86	0.06
Specificity	0.51	0.13	0.52	0.13	0.59	0.12	0.53	0.13	0.53	0.12
NPV	0.72	0.14	0.65	0.13	0.71	0.13	0.71	0.13	0.77	0.13

Table 6.3: Scenario 1 of Emotional Imagery Phase Dataset

Accuracy was consistent among all algorithms, ranging from **0.80 to 0.81**, indicating that all models achieved similar rates of overall correct predictions. The **F1-score**, which balances recall and precision, was also largely consistent, with most models achieving approximately **0.72 ± 0.07**. Notably, the **Bagging Decision Tree** and **Extra Trees** models showed a slight improvement, reaching **0.75 ± 0.07**, reflecting a marginally better balance between sensitivity and precision.

In terms of **recall (sensitivity)**, which is crucial for identifying as many truly stressed patients as possible, the **Extra Trees classifier** achieved the highest score (**0.93 ± 0.05**), followed closely by **Gradient Boosting Decision Tree** with **0.91 ± 0.05**. These results indicate that both models were particularly effective at minimizing false negatives.

Precision, which reflects the proportion of correctly identified stressed patients among those predicted as stressed, was highest for the **Bagging Decision Tree** at **0.84 ± 0.04**, while the other models produced values in the range of **0.82 to 0.83**, suggesting relatively low rates of false positives.

Regarding the **Area Under the Curve (AUC)** metric, which evaluates the classifier's ability to distinguish between classes, the **Adaptive Boosting Decision Tree** exhibited the lowest performance at **0.80 ± 0.08**, whereas the remaining models fell between **0.84 and 0.86**, indicating satisfactory overall discrimination.

One of the more challenging aspects across all models was **specificity**, or the true negative rate. The highest value was recorded by the **Bagging Decision Tree** (**0.59 ± 0.12**), while the others remained within a lower range of **0.51 to 0.53**, suggesting a tendency for some non-stressed individuals to be incorrectly classified as stressed.

Finally, for **Negative Predictive Value (NPV)**, the **Extra Trees** model again performed best, with a score of **0.77 ± 0.13**, meaning it was more effective in correctly identifying non-stressed patients. In contrast, the **Adaptive Boosting Decision Tree** showed the lowest NPV at **0.65 ± 0.13**, indicating weaker performance in this aspect.

Overall, based on these original results, the **Extra Trees classifier** stood out as the most effective model in this scenario. Its superior recall, NPV, and balanced performance across key metrics underscore its suitability for reliably detecting high-stress patients during the emotional imagery experiment. This is particularly important in the context of clinical support, where identifying vulnerable individuals accurately and minimizing both false positives and false negatives is essential.

6.2.2 Scenario 2 of Emotional Imagery Phase Dataset

In **Scenario 2**, which focuses on classifying patients into high and low stress categories based on the **median DASS score (threshold = 5)**, the original classification results showed more moderate overall performance compared to Scenario 1, as illustrated in Table 6.4.

	Gradient Boosting Decision Tree		Adaptive Boosting Decision Tree		Bagging Decision Tree		Random Forests		Extra Trees	
	Value	**SD**	**Value**	**SD**	**Value**	**SD**	**Value**	**SD**	**Value**	**SD**
Accuracy	0.68	0.07	0.64	0.08	0.67	0.07	0.67	0.07	0.64	0.07
F1-score	0.68	0.07	0.63	0.09	0.66	0.07	0.66	0.07	0.63	0.07
Recall-sensitivity	0.70	0.09	0.67	0.10	0.68	0.07	0.69	0.09	0.68	0.09
Precision (PPV)	0.72	0.08	0.67	0.09	0.70	0.07	0.70	0.07	0.66	0.06
AUC	0.76	0.07	0.70	0.08	0.74	0.07	0.73	0.07	0.71	0.07
Specificity	0.66	0.12	0.60	0.13	0.65	0.11	0.64	0.10	0.59	0.10
NPV	0.65	0.08	0.61	0.09	0.63	0.07	0.64	0.08	0.61	0.08

Table 6.4: Scenario 2 of Emotional Imagery Phase Dataset

Accuracy values were relatively low across all models. The **Gradient Boosting Decision Tree** achieved the highest accuracy (**0.68 ± 0.07**), while **Adaptive Boosting Decision Tree** and **Extra Trees** recorded the lowest (**0.64 ± 0.08**). The remaining models showed nearly identical values around **0.67 ± 0.07**, indicating minor variability in the models' overall correctness.

In terms of the **F1-score**, which balances recall and precision, **Gradient Boosting Decision Tree** again outperformed the other algorithms (**0.68 ± 0.07**), followed by a group of models with similar values (**0.66 ± 0.07**). The lowest F1-scores were observed in **Adaptive Boosting Decision Tree** and **Extra Trees**, both with **0.63 ± 0.07**, suggesting these models were slightly less effective at achieving a balance between true positive predictions and false positives.

Recall (sensitivity), a key metric for detecting stressed patients, ranged between **0.67 and 0.70** across all models, indicating consistent performance in identifying positive cases. While no model significantly outperformed others in this regard, the uniformity in scores suggests that all models were moderately capable of identifying stressed individuals based on the median threshold.

Precision (PPV) revealed more variation. The **Gradient Boosting Decision Tree** achieved the highest precision (**0.72 ± 0.08**), indicating strong performance in minimizing false positives. In contrast, **Extra Trees** recorded the lowest precision (**0.66 ± 0.06**), with **Adaptive Boosting Decision Tree** slightly higher at **0.67 ± 0.09**. The remaining algorithms produced similar precision values of approximately **0.70 ± 0.07**.

Regarding **AUC (Area Under the Curve)**, which evaluates the model's ability to discriminate between classes, **Gradient Boosting Decision Tree** again had the best result (**0.76 ± 0.07**), while **Adaptive Boosting Decision Tree** showed the lowest (**0.70 ± 0.08**). Other models ranged between **0.71 and 0.74**, reflecting moderate discriminative performance.

Specificity, which measures the ability to correctly identify non-stressed individuals, was generally lower across the board. The **Extra Trees classifier** had the weakest performance (**0.59 ± 0.10**), followed by **AdaBoost** (**0.60 ± 0.13**). The remaining models fell between **0.64 and 0.66**, showing slightly improved performance in recognizing the negative class.

Lastly, for **Negative Predictive Value (NPV)**, **Extra Trees** and **AdaBoost** again recorded the lowest scores (**0.61 ± 0.08**), while other models ranged from **0.63 to 0.65**. These results indicate that the models had limited reliability in identifying non-stressed patients when using the median DASS cut-off.

In summary, under the original settings of Scenario 2, the **Gradient Boosting Decision Tree** algorithm consistently achieved the best overall performance across nearly all evaluation metrics, including accuracy, F1-score, precision, and AUC. This suggests it is the most appropriate choice for classifying patients based on a **balanced DASS threshold**. However, the overall results were weaker compared to Scenario 1, likely due to the **more challenging distribution of samples** caused by using the **median value** as the threshold, which creates more balanced but less separable class distributions. These findings highlight the need for further strategies, such as data balancing techniques, to improve the model's ability to generalize, particularly in complex, real-world datasets.

6.2.3 Scenario 3 of Emotional Imagery Phase Dataset

In **Scenario 3**, where the classification task aimed to distinguish between high and low psychological flexibility based on an **AAQ-II threshold of 24**, the original results (prior to any oversampling technique) highlighted notable differences in model performance. These original results are as illustrated in Table 6.5.

	Gradient Boosting Decision Tree		Adaptive Boosting Decision Tree		Bagging Decision Tree		Random Forests		Extra Trees	
	Value	**SD**	**Value**	**SD**	**Value**	**SD**	**Value**	**SD**	**Value**	**SD**
Accuracy	0.84	0.04	0.82	0.05	0.85	0.04	0.85	0.05	0.85	0.05
F1-score	0.72	0.08	0.67	0.09	0.75	0.08	0.74	0.08	0.74	0.08
Recall-sensitivity	0.48	0.14	0.38	0.15	0.53	0.16	0.51	0.16	0.50	0.15
Precision (PPV)	0.68	0.17	0.61	0.21	0.68	0.17	0.70	0.18	0.69	0.17
AUC	0.82	0.08	0.77	0.09	0.83	0.08	0.83	0.08	0.82	0.08
Specificity	0.94	0.04	0.93	0.05	0.93	0.04	0.94	0.04	0.94	0.04
NPV	0.87	0.03	0.85	0.03	0.89	0.03	0.88	0.04	0.88	0.03

Table 6.5: Scenario 3 of Emotional Imagery Phase Dataset

Accuracy scores ranged from **0.82 ± 0.05** (Adaptive Boosting Decision Tree – the lowest) to values between **0.84 and 0.85** for the remaining algorithms, indicating solid overall classification ability across models, except for Adaptive Boosting Decision Tree, which underperformed.

The **F1-score**, which reflects the harmonic mean of precision and recall, was also lowest for **Adaptive Boosting Decision Tree** (**0.67 ± 0.09**), while the other algorithms achieved more favorable results ranging from **0.72 to 0.75**, suggesting a more balanced predictive capacity in those models.

Recall (sensitivity), a particularly important metric for identifying patients with **low psychological flexibility** (i.e., those more likely to avoid or disengage from unpleasant experiences), was weakest in **Adaptive Boosting Decision Tree** (**0.38 ± 0.15**), indicating a high rate of false negatives. The other algorithms ranged between **0.48 and 0.53**, which, while still moderate, demonstrated better ability to correctly identify at-risk individuals.

In terms of **Precision (PPV)**, **Adaptive Boosting Decision Tree** again showed the lowest performance (**0.61 ± 0.21**), whereas the remaining classifiers produced scores between **0.68 and 0.70**, suggesting more reliable identification of truly inflexible patients among those predicted as such.

AUC (Area Under the Curve) values followed a similar pattern: **Adaptive Boosting Decision Tree** yielded the lowest score (**0.77 ± 0.09**), while all other algorithms ranged from **0.82 to 0.83**, indicating strong discriminative power between high and low psychological flexibility classes.

Interestingly, **Specificity** was consistently high across all models, with values between **0.93 and 0.94**, showing that all classifiers were highly effective at correctly identifying patients with **high psychological flexibility** (i.e., those who demonstrate acceptance and psychological resilience).

Finally, for **Negative Predictive Value (NPV)**, a key metric for evaluating how accurately non-inflexible patients are identified, **Adaptive Boosting Decision Tree** again performed the worst (**0.85 ± 0.03**), while the remaining classifiers achieved values between **0.87 and 0.89**.

Overall, the results of Scenario 3 show **strong specificity and accuracy**, meaning that most patients with **high psychological flexibility** were correctly identified. However, the **recall values were relatively low across all models**, particularly for **Adaptive Boosting Decision Tree**, suggesting that patients with **low psychological flexibility** (those most in need of psychological support) were frequently missed.

Among the models, **Random Forest, Extra Trees, and Gradient Boosting Decision Tree** demonstrated more favorable and balanced performance across most metrics, especially in **F1-score, AUC, precision**, and **NPV**, making them more reliable choices for the task of identifying cancer patients with **reduced psychological flexibility** during the emotional imagery experiment.

These findings highlight the challenge of this classification task, especially due to the imbalance in how easily flexible versus inflexible patients can be identified using physiological signals alone. The results suggest the need for enhancement strategies to better capture this psychologically important but more subtle trait in future iterations.

6.2.4 Scenario 4 of Emotional Imagery Phase Dataset

In **Scenario 4**, which focuses on classifying patients according to their **psychological flexibility** based on the **median AAQ-II score (13.5)**, the original results, before applying any data balancing techniques, reveal moderate but consistent differences in classifier performance as illustrated in Table 6.6.

	Gradient Boosting Decision Tree		Adaptive Boosting Decision Tree		Bagging Decision Tree		Random Forests		Extra Trees	
	Value	**SD**	**Value**	**SD**	**Value**	**SD**	**Value**	**SD**	**Value**	**SD**
Accuracy	0.73	0.06	0.69	0.06	0.76	0.06	0.76	0.06	0.77	0.07
F1-score	0.72	0.06	0.69	0.06	0.76	0.06	0.76	0.06	0.77	0.07
Recall-sensitivity	0.72	0.10	0.70	0.09	0.75	0.10	0.75	0.09	0.74	0.10
Precision (PPV)	0.73	0.07	0.69	0.08	0.77	0.07	0.77	0.08	0.79	0.08
AUC	0.81	0.05	0.76	0.06	0.82	0.06	0.83	0.05	0.84	0.05
Specificity	0.73	0.10	0.68	0.10	0.77	0.09	0.78	0.09	0.79	0.09
NPV	0.73	0.07	0.70	0.07	0.76	0.07	0.76	0.07	0.76	0.08

Table 6.6: Scenario 4 of Emotional Imagery Phase Dataset

Accuracy scores were highest among most classifiers, ranging between **0.73 and 0.77**, while **Adaptive Boosting Decision Tree** recorded the lowest accuracy at **0.69 ± 0.06**, suggesting a less effective general classification ability in this context.

Regarding the **F1-score**, which reflects the balance between precision and recall, most models performed adequately with values between **0.72 and 0.77**. In contrast, **Adaptive Boosting Decision Tree** again underperformed, achieving the lowest score at **0.69 ± 0.06**, indicating reduced reliability in handling the trade-off between false positives and false negatives.

Recall (sensitivity), which measures the classifier's ability to correctly detect individuals with **low psychological flexibility**, a trait often associated with avoidance coping, was lowest for **Adaptive Boosting Decision Tree** (**0.70 ± 0.09**), while the rest of the algorithms achieved better results in the range of **0.72 to 0.75**.

In terms of **Precision (PPV)**, **Adaptive Boosting Decision Tree** again yielded the weakest performance (**0.69 ± 0.08**), while the remaining models showed values between **0.73 and 0.79**, meaning they were more effective in correctly identifying low-flexibility patients among all those predicted as such.

AUC (Area Under the Curve) values, reflecting the ability to discriminate between high and low psychological flexibility, were highest for the better-performing models, ranging between **0.81 and 0.84**. **Adaptive Boosting Decision Tree** recorded the lowest AUC at **0.76 ± 0.06**, indicating limited class separation performance.

The **Specificity** scores, which represent the true negative rate (correct identification of high-flexibility patients), ranged from **0.73 to 0.79** for most algorithms, while **Adaptive Boosting Decision Tree** again lagged with a score of **0.68 ± 0.10**, indicating a higher false positive rate.

Finally, **Negative Predictive Value (NPV)** followed the same trend, with **Adaptive Boosting Decision Tree** reporting the lowest (**0.70 ± 0.07**), while other models achieved results between **0.73 and 0.76**, suggesting a more reliable identification of high-flexibility individuals.

The findings from Scenario 4 demonstrate that most classifiers—especially **Gradient Boosting Decision Tree**, **Random Forest**, and **Extra Trees**—performed reasonably well across key metrics when attempting to detect differences in psychological flexibility using the **median AAQ-II threshold**. This scenario, compared to Scenario 3, benefits from a more balanced distribution of samples due to the use of the median split, which supports slightly higher performance scores overall.

Once again, **Adaptive Boosting Decision Tree** consistently performed the worst across all evaluated metrics, showing limitations in both sensitivity and precision, and resulting in lower F1 and AUC values. This underlines its reduced suitability for detecting patients with **low psychological flexibility** in this dataset.

Models like **Extra Trees and Gradient Boosting Decision Tree**, which achieved strong F1-scores, precision, and AUC, appear better suited for this classification task. Their ability to simultaneously minimize false positives and false negatives suggests a more balanced and stable classification behavior.

These results are particularly relevant for the goals of this thesis, as accurately distinguishing between **patients with adaptive and maladaptive psychological responses** can inform more tailored intervention strategies in the context of cancer care and Acceptance and Commitment Therapy.

6.2.5 Scenario of EMA Phase Dataset

In the **Scenario of the EMA phase dataset**, the original classification results demonstrated moderate to low overall performance across the evaluated machine learning models, with particular challenges observed in recall (sensitivity) and F1-score values. These scores are illustrated in Table 6.7.

	Gradient Boosting Decision Tree		Ada Boosting Decision Tree		Bagging Decision Tree		Random Forests		Extra Trees	
	Value	**SD**	**Value**	**SD**	**Value**	**SD**	**Value**	**SD**	**Value**	**SD**
Accuracy	0.72	0.12	0.70	0.11	0.75	0.11	0.74	0.10	0.78	0.11
F1-score	0.60	0.17	0.58	0.15	0.64	0.17	0.61	0.16	0.66	0.17
Recall-sensitivity	0.34	0.28	0.35	0.25	0.38	0.27	0.31	0.23	0.40	0.28
Precision (PPV)	nan	nan	nan	nan	nan	nan	nan	nan	nan	nan
AUC	0.70	0.18	0.64	0.18	0.71	0.18	0.71	0.16	0.73	0.18
Specificity	0.88	0.12	0.84	0.14	0.90	0.12	0.92	0.11	0.93	0.10
NPV	0.77	0.10	0.76	0.08	0.79	0.09	0.77	0.07	0.80	0.09

Table 6.7: Scenario of EMA Phase Dataset

Accuracy scores showed noticeable variation across models. The **Extra Trees** classifier achieved the highest accuracy (0.78 ± 0.11), followed closely by **Bagging Decision Tree** (0.75 ± 0.11) and **Random Forests** (0.74 ± 0.10). The **Gradient Boosting Decision Tree** and **Adaptive Boosting Decision Tree** models presented the lowest accuracy scores, at 0.72 ± 0.12 and 0.70 ± 0.11 respectively, indicating that ensemble-based methods like Bagging and Extra Trees were better suited for this dataset in terms of overall correctness.

In terms of **F1-score**, which balances precision and recall, **Extra Trees** again delivered the best performance (0.66 ± 0.17), closely followed by **Bagging Decision Tree** (0.64 ± 0.17). The **Gradient Boosting Decision Tree** and **Random Forest** models produced moderate F1-scores (0.60 ± 0.17 and 0.61 ± 0.16, respectively), while **Adaptive Boosting Decision Tree** exhibited the lowest value (0.58 ± 0.15). This suggests that models like Extra Trees were better able to maintain a balance between sensitivity and precision under the conditions of the EMA dataset.

Regarding **Recall (Sensitivity)**, which is crucial for identifying acceptance or avoidance states in patients, the results were overall quite low due to **data imbalance** as shown in Table 6.2 . The **Extra Trees** classifier achieved the highest recall (0.40 ± 0.28), followed by **Bagging Decision Tree** (0.38 ± 0.27). **Gradient Boosting Decision Tree** and **Adaptive Boosting Decision Tree** presented lower recall values of 0.34 ± 0.28 and 0.35 ± 0.25, respectively, while **Random Forests** showed the weakest recall (0.31 ± 0.23). These findings highlight the difficulty all models faced in accurately identifying minority class samples in this particular dataset.

Precision (PPV) values were not available (NaN) across all models in this scenario. This likely reflects the high imbalance or label distribution issues in the original dataset, causing instability in precision calculations.

For **AUC (Area Under the Curve)**, which reflects the models' ability to discriminate between classes, **Extra Trees** led again (0.73 ± 0.18), followed by **Random Forests** and **Bagging Decision Tree** (both around 0.71 ± 0.16–0.18). **Gradient Boosting Decision Tree** showed slightly lower AUC (0.70 ± 0.18), and **Adaptive Boosting Decision Tree** recorded the lowest (0.64 ± 0.18). Overall, the discriminative ability of the models was moderate, with Extra Trees being marginally more reliable.

Specificity, which measures the correct identification of patients classified as acceptance, showed stronger results compared to sensitivity. **Extra Trees** achieved the highest specificity (0.93 ± 0.10), indicating strong ability to correctly classify the negative class.

Random Forests and **Bagging Decision Tree** also performed well, achieving specificity scores of 0.92 ± 0.11 and 0.90 ± 0.12 respectively, while **Gradient Boosting Decision Tree** and **Adaptive Boosting Decision Tree** scored slightly lower (0.88 ± 0.12 and 0.84 ± 0.14).

Finally, in terms of **Negative Predictive Value (NPV)**, **Extra Trees** again outperformed other models (0.80 ± 0.09), followed by **Bagging Decision Tree** (0.79 ± 0.09) and **Random Forests** (0.77 ± 0.07). **Gradient Boosting Decision Tree** and **Adaptive Boosting Decision Tree** showed slightly lower NPV values (0.77 ± 0.10 and 0.76 ± 0.08 respectively), confirming the superior performance of ensemble-based classifiers for this dataset.

Overall, the results indicate that **Extra Trees** consistently performed best across most evaluation metrics in the EMA dataset scenario, highlighting its robustness in handling complex and imbalanced clinical data regarding patients' acceptance versus avoidance categorization.

6.3 Oversampling Techniques

Oversampling is a widely used data preprocessing strategy in supervised machine learning, particularly when dealing with **imbalanced datasets**, a common occurrence where one class is significantly underrepresented compared to others. In such cases, machine learning models tend to be biased toward the majority class, leading to poor classification performance for the minority class, which often represents the more critical or high-risk category [48].

To mitigate this issue, oversampling techniques artificially increase the number of samples in the minority class, creating a more balanced dataset and helping the classifier to better learn the underlying patterns of both classes. This process enhances the model's ability to generalize, particularly in scenarios where identifying the minority class is essential, such as in the detection of **high stress**, **low psychological flexibility**, or **avoidance coping** in cancer patients, core objectives of this thesis.

Various oversampling methods exist to address this problem, with some of the most applied being **Random Oversampling**, **SMOTE (Synthetic Minority Over-sampling Technique)**, and **ADASYN (Adaptive Synthetic Sampling)**. These techniques are further described in Sections 6.3.1 to 6.3.3.

6.3.1 Random Oversampling

Random Oversampling is one of the simplest and most used techniques for addressing class imbalance in supervised learning tasks. The core idea behind this method is to balance the dataset by **randomly duplicating instances from the minority class** until both classes have approximately the same number of samples [48]. This straightforward approach ensures that the learning algorithm receives sufficient exposure to minority class examples during training, which helps reduce the bias toward the majority class and improves the model's ability to detect rare but important cases. Random Oversampling has been shown to be effective in improving recall and reducing classification errors for the minority class.

6.3.2 SMOTE (Synthetic Minority Over-sampling Technique)

SMOTE (Synthetic Minority Over-sampling Technique) is one of the most widely adopted oversampling methods for addressing class imbalance in classification problems. Unlike random oversampling, this algorithm works by selecting a random sample from the minority class and identifying its k nearest minority neighbors. It then selects one of these neighbors at random and creates a synthetic sample along the line segment connecting the two instances. This interpolation-based strategy ensures that the new samples are close to existing ones in the feature space but introduce enough variation to support generalization during training [48].

6.3.3 ADASYN (Adaptive Synthetic Sampling)

ADASYN (Adaptive Synthetic Sampling Approach for Imbalanced Learning) builds upon the principles of SMOTE by not only generating synthetic minority class samples but also by focusing on regions of the feature space where the minority class is underrepresented or harder to learn [49]. The core idea of ADASYN is to adaptively determine the number of synthetic samples to generate for each minority class instance based on its **local data distribution**. Specifically, more synthetic data is created for those samples that are surrounded by a large number of majority class neighbors, for example the samples that lie in **complex or overlapping regions** of the decision space. Conversely, fewer or no synthetic instances are created for easier-to-learn samples in well-separated areas. This adaptive behavior enables the classifier to **focus on difficult decision boundaries**, thereby improving its sensitivity and generalization in imbalanced contexts [48].

6.4 Classifier Selection after Oversampling Application

In this section, the results after the application of the three oversampling techniques explained in Sections 6.3.1, 6.3.2, and 6.3.3 are presented. It is important to clarify why **Random Oversampling** and **SMOTE** were successfully implemented across all scenarios, while **ADASYN** was not applicable in Scenarios 2 and 4.

Random Oversampling and SMOTE are techniques that operate in a mechanical manner, aiming to balance class distributions by either duplicating minority class samples (Random Oversampling) or generating synthetic samples through interpolation (SMOTE), regardless of the initial level of balance in the dataset. Consequently, even in scenarios where the classes were already relatively balanced, such as Scenario 2 (DASS median threshold) and Scenario 4 (AAQII median threshold), these methods continued to function and artificially balanced the datasets further.

In contrast, ADASYN adopts an adaptive behavior based on the local data structure. It generates synthetic samples primarily in regions where the minority class is underrepresented and where the classification difficulty is higher. When the minority and majority classes are already well-balanced, and the minority samples are not located in complex or ambiguous regions, ADASYN naturally determines that there is no significant need for additional synthetic samples. As a result, in Scenarios 2 and 4, ADASYN did not generate any new samples and, therefore, was not further applied. This behavior highlights the algorithm’s sensitivity to local density distributions and its effort to avoid introducing noise in well-separated class regions. By prioritizing sample generation only in challenging zones, ADASYN helps maintain class separability and reduce the risk of overfitting.

6.4.1 Scenario 1 of Emotional Imagery Phase Dataset after Oversampling

This section presents the classification results for Scenario 1 of the emotional imagery phase dataset after the application of three different oversampling techniques: Random Oversampling, SMOTE, and ADASYN. To evaluate the performance, five machine learning algorithms were used: Gradient Boosting Decision Tree, Adaptive Boosting Decision Tree, Bagging Decision Tree, Random Forests, and Extra Trees. The performance metrics include Accuracy, F1-score, Recall (Sensitivity), Precision (Positive Predictive Value), Area Under the Curve (AUC), Specificity, and Negative Predictive Value (NPV), along with their corresponding standard deviations (SD). The following tables (Tables 6.8–6.10) summarize the results obtained after applying each oversampling method.

	Gradient Boosting Decision Tree		Adaptive Boosting Decision Tree		Bagging Decision Tree		Random Forests		Extra Trees	
	Value	**SD**	**Value**	**SD**	**Value**	**SD**	**Value**	**SD**	**Value**	**SD**
Accuracy	0.86	0.04	0.82	0.05	0.89	0.04	0.91	0.04	0.93	0.03
F1-score	0.86	0.04	0.82	0.05	0.89	0.04	0.91	0.04	0.93	0.03
Recall-sensitivity	0.8	0.07	0.77	0.08	0.82	0.07	0.85	0.07	0.92	0.05
Precision(PPV)	0.92	0.05	0.85	0.06	0.96	0.04	0.96	0.04	0.95	0.04
AUC	0.94	0.03	0.89	0.04	0.97	0.02	0.98	0.02	0.98	0.01
Specificity	0.93	0.05	0.86	0.07	0.96	0.03	0.97	0.03	0.95	0.04
NPV	0.83	0.05	0.79	0.06	0.85	0.05	0.87	0.05	0.92	0.05

Table 6.8: Results after Random Oversampling

	Gradient Boosting Decision Tree		Adaptive Boosting Decision Tree		Bagging Decision Tree		Random Forests		Extra Trees	
	Value	**SD**	**Value**	**SD**	**Value**	**SD**	**Value**	**SD**	**Value**	**SD**
Accuracy	0.81	0.05	0.79	0.05	0.83	0.05	0.83	0.05	0.85	0.04
F1-score	0.81	0.05	0.79	0.05	0.82	0.05	0.83	0.05	0.85	0.04
Recall-sensitivity	0.77	0.07	0.76	0.08	0.78	0.08	0.79	0.07	0.82	0.07
Precision(PPV)	0.84	0.06	0.81	0.06	0.86	0.06	0.86	0.06	0.87	0.05
AUC	0.89	0.04	0.86	0.05	0.91	0.04	0.91	0.04	0.93	0.03
Specificity	0.85	0.06	0.82	0.07	0.87	0.06	0.87	0.06	0.88	0.06
NPV	0.79	0.05	0.77	0.05	0.80	0.06	0.81	0.06	0.83	0.05

Table 6.9: Results after SMOTE

	Gradient Boosting Decision Tree		Adaptive Boosting Decision Tree		Bagging Decision Tree		Random Forests		Extra Trees	
	Value	**SD**	**Value**	**SD**	**Value**	**SD**	**Value**	**SD**	**Value**	**SD**
Accuracy	0.78	0.05	0.75	0.05	0.82	0.05	0.81	0.05	0.82	0.05
F1-score	0.78	0.05	0.75	0.05	0.81	0.05	0.81	0.05	0.82	0.05
Recall-sensitivity	0.71	0.08	0.7	0.08	0.76	0.07	0.76	0.08	0.78	0.07
Precision(PPV)	0.83	0.06	0.78	0.06	0.86	0.06	0.85	0.07	0.85	0.06
AUC	0.86	0.05	0.81	0.05	0.89	0.04	0.89	0.04	0.91	0.04
Specificity	0.85	0.06	0.8	0.07	0.87	0.06	0.86	0.07	0.86	0.06
NPV	0.75	0.06	0.73	0.06	0.79	0.06	0.78	0.06	0.79	0.05

Table 6.10: Results after ADASYN

After the application of the three oversampling techniques, significant improvements were observed in the classification results for Scenario 1. Compared to the initial imbalanced dataset (Table 6.3), where classifiers showed limited performance, particularly in terms of precision, F1-score, and specificity, the optimized models achieved notable enhancements across all key metrics.

Among the oversampling methods, **Random Oversampler** delivered the **best overall performance**, recording the **highest recall (92%)**, **precision (95%)**, **F1-score (93%)**, and **AUC (98%)**, establishing it as the most effective technique for this scenario. These figures mark substantial gains when contrasted with the original scenario, where average recall was already relatively high (~93% for Extra Trees), but **precision** and **F1-score** were considerably lower, averaging around **83%** and **75%**, respectively. Furthermore, **specificity**, which was particularly low in the initial setup (around 51–53%), improved dramatically to over **94%** under Random Oversampling, indicating a significant reduction in false positives.

The high recall value demonstrates that the majority of stressed patients were correctly identified, effectively minimizing false negatives, an essential factor in applications involving the early detection of stress in cancer patients. In parallel, the high precision value ensures that very few non-stressed individuals were misclassified as stressed, thus supporting accurate risk stratification and minimizing unnecessary interventions. The exceptional AUC score of **98%**, up from **~0.84** in the original scenario, further confirms the model's enhanced ability to distinguish between stressed and non-stressed cases after addressing class imbalance.

Regarding the machine learning algorithms, the **Extra Trees classifier** consistently outperformed the others after oversampling, achieving the **highest recall, precision, F1-score, and AUC values** across all three oversampling strategies. In the baseline (non-oversampled) scenario, Extra Trees had strong recall (93%) but was constrained by a moderate F1-score (75%) and relatively low specificity (53%). With Random Oversampling, however, it achieved near-optimal results across all metrics, making it the most effective classifier for Scenario 1.

Although **SMOTE** and **ADASYN** also improved performance compared to the original scenario—particularly in reducing false positives and increasing AUC—their overall impact was slightly less pronounced. **SMOTE generally outperformed ADASYN**, but **neither matched the consistent and high-performing results** obtained with Random Oversampling.

6.4.2 Scenario 2 of Emotional Imagery Phase Dataset after Oversampling

This section presents the classification results for Scenario 2 of the emotional imagery phase dataset, specifically using the DASS with a median threshold of five (5), after the application of Random Oversampling and SMOTE techniques. The same five machine-learning algorithms were evaluated: Gradient Boosting Decision Tree, Adaptive Boosting Decision Tree, Bagging Decision Tree, Random Forests, and Extra Trees. Performance metrics reported include Accuracy, F1-score, Recall (Sensitivity), Precision (Positive Predictive Value), Area Under the Curve (AUC), Specificity, and Negative Predictive Value (NPV), along with their corresponding standard deviations (SD). The following tables (Tables 6.11–6.12) summarize the classification results obtained after applying each oversampling method.

	Gradient Boosting Decision Tree		Adaptive Boosting Decision Tree		Bagging Decision Tree		Random Forests		Extra Trees	
	Value	**SD**	**Value**	**SD**	**Value**	**SD**	**Value**	**SD**	**Value**	**SD**
Accuracy	0.71	0.06	0.65	0.06	0.71	0.06	0.72	0.06	0.69	0.06
F1-score	0.71	0.06	0.64	0.06	0.71	0.06	0.71	0.07	0.69	0.06
Recall-sensitivity	0.64	0.10	0.59	0.11	0.66	0.10	0.67	0.10	0.67	0.10
Precision(PPV)	0.75	0.07	0.67	0.07	0.74	0.07	0.74	0.07	0.70	0.07
AUC	0.79	0.06	0.71	0.07	0.79	0.06	0.79	0.06	0.80	0.06
Specificity	0.78	0.08	0.71	0.08	0.76	0.08	0.76	0.08	0.71	0.09
NPV	0.69	0.06	0.64	0.06	0.69	0.06	0.70	0.07	0.69	0.07

Table 6.11: Results after Random Oversampling

	Gradient Boosting Decision Tree		Adaptive Boosting Decision Tree		Bagging Decision Tree		Random Forests		Extra Trees	
	Value	**SD**	**Value**	**SD**	**Value**	**SD**	**Value**	**SD**	**Value**	**SD**
Accuracy	0.7	0.07	0.65	0.07	0.69	0.06	0.7	0.07	0.67	0.07
F1-score	0.7	0.07	0.65	0.07	0.69	0.07	0.69	0.07	0.67	0.07
Recall-sensitivity	0.65	0.1	0.6	0.11	0.66	0.09	0.66	0.1	0.64	0.1
Precision(PPV)	0.72	0.08	0.67	0.09	0.71	0.08	0.71	0.08	0.68	0.08
AUC	0.78	0.06	0.72	0.08	0.77	0.06	0.77	0.06	0.75	0.06
Specificity	0.74	0.08	0.71	0.08	0.72	0.09	0.73	0.09	0.70	0.09
NPV	0.69	0.07	0.65	0.07	0.68	0.07	0.68	0.07	0.66	0.07

Table 6.12: Results after SMOTE

After the application of the three oversampling techniques, the classification performance for Scenario 2 (DASS with Median Threshold 5) demonstrated noticeable improvements across several evaluation metrics. Compared to the initial results without oversampling (Table 6.4), where the models exhibited moderate performance, the optimized results show enhanced recall, F1-score, and AUC values, indicating improved sensitivity and discriminative ability of the models.

Random Oversampler and SMOTE produced very similar outcomes, both achieving recall values of approximately 71%, precision around 72%, and F1-scores close to 71%. These figures reflect a more balanced classification capability compared to the baseline.

The AUC was slightly higher for Random Oversampling, reaching up to 0.80 in the case of Extra Trees and 0.79 for Bagging and Random Forests, while in the initial scenario, AUC values ranged from 0.70 to 0.76. This improvement suggests that the models became more reliable in distinguishing between stressed and non-stressed individuals after oversampling.

When comparing to the baseline results, the F1-score improved from an initial range of 63–68% to approximately 71% after oversampling. Recall values, which initially ranged from 67% to 70%, also increased to around 71%, reflecting a better ability to detect true stressed cases. Precision remained relatively stable, close to 70–72%, indicating that false positives did not increase significantly despite gains in sensitivity. The modest increase in AUC—from around 0.73–0.76 to 0.79–0.80—confirms a consistent, although not dramatic, improvement in overall classification performance.

Despite these gains, the overall performance of the models in Scenario 2 remained slightly lower than that observed in Scenario 1, likely due to the increased classification difficulty introduced by the more balanced label distribution and the specific DASS threshold applied. Nevertheless, these results demonstrate that oversampling techniques effectively improved model robustness in a more challenging classification context.

Among the machine learning algorithms evaluated, the Bagging Decision Tree achieved the highest recall and AUC values following the application of oversampling methods. This indicates that Bagging was the most effective classifier for identifying stressed patients in this scenario, while maintaining competitive results in other evaluation metrics. While results from ADASYN were not included in this analysis, the trends observed with Random Oversampling and SMOTE suggest that traditional oversampling techniques contributed positively to classification performance without compromising the model's precision.

6.4.3 Scenario 3 of Emotional Imagery Phase Dataset after Oversampling

This section presents the classification results for Scenario 3 of the emotional imagery phase dataset, specifically based on the AAQ-II with a threshold of 24, following the application of three oversampling techniques: Random Oversampling, SMOTE, and ADASYN. As in the previous scenarios, the machine learning algorithms evaluated include Gradient Boosting Decision Tree, Adaptive Boosting Decision Tree, Bagging

Decision Tree, Random Forests, and Extra Trees. Performance metrics such as Accuracy, F1-score, Recall (Sensitivity), Precision (Positive Predictive Value), Area Under the Curve (AUC), Specificity, and Negative Predictive Value (NPV) were assessed and reported with their respective standard deviations (SD). The results obtained after applying each oversampling method are summarized in Tables 6.13–6.15.

	Gradient Boosting Decision Tree		Adaptive Boosting Decision Tree		Bagging Decision Tree		Random Forests		Extra Trees	
	Value	**SD**	**Value**	**SD**	**Value**	**SD**	**Value**	**SD**	**Value**	**SD**
Accuracy	0.89	0.04	0.83	0.04	0.94	0.03	0.94	0.03	0.96	0.03
F1-score	0.89	0.04	0.83	0.04	0.94	0.03	0.94	0.03	0.96	0.03
Recall-sensitivity	0.93	0.05	0.88	0.07	0.99	0.02	0.99	0.02	0.98	0.02
Precision(PPV)	0.87	0.05	0.81	0.05	0.91	0.05	0.91	0.04	0.93	0.04
AUC	0.95	0.03	0.91	0.04	0.98	0.02	0.98	0.02	0.99	0.01
Specificity	0.86	0.06	0.79	0.06	0.89	0.06	0.90	0.05	0.93	0.04
NPV	0.92	0.07	0.87	0.06	0.99	0.03	0.98	0.02	0.98	0.02

Table 6.13: Results after random oversampling

	Gradient Boosting Decision Tree		Adaptive Boosting Decision Tree		Bagging Decision Tree		Random Forests		Extra Trees	
	Value	**SD**	**Value**	**SD**	**Value**	**SD**	**Value**	**SD**	**Value**	**SD**
Accuracy	0.85	0.04	0.78	0.05	0.88	0.04	0.88	0.03	0.89	0.04
F1-score	0.85	0.04	0.78	0.05	0.88	0.04	0.88	0.03	0.89	0.04
Recall-sensitivity	0.86	0.06	0.81	0.07	0.90	0.06	0.89	0.05	0.91	0.05
Precision(PPV)	0.84	0.05	0.77	0.05	0.87	0.05	0.88	0.05	0.88	0.05
AUC	0.92	0.03	0.86	0.04	0.94	0.02	0.94	0.02	0.94	0.02
Specificity	0.83	0.07	0.76	0.07	0.87	0.06	0.87	0.05	0.87	0.06
NPV	0.86	0.05	0.80	0.06	0.90	0.05	0.89	0.05	0.90	0.05

Table 6.14: Results after SMOTE

	Gradient Boosting Decision Tree		Adaptive Boosting Decision Tree		Bagging Decision Tree		Random Forests		Extra Trees	
	Value	**SD**	**Value**	**SD**	**Value**	**SD**	**Value**	**SD**	**Value**	**SD**
Accuracy	0.82	0.05	0.77	0.05	0.87	0.04	0.87	0.04	0.87	0.04
F1-score	0.82	0.05	0.77	0.05	0.87	0.04	0.87	0.04	0.87	0.04
Recall-sensitivity	0.9	0.05	0.81	0.07	0.91	0.05	0.91	0.05	0.91	0.05
Precision(PPV)	0.78	0.05	0.75	0.05	0.84	0.05	0.84	0.05	0.85	0.05
AUC	0.89	0.04	0.83	0.05	0.93	0.03	0.93	0.03	0.94	0.03
Specificity	0.75	0.08	0.73	0.07	0.82	0.07	0.82	0.06	0.84	0.06
NPV	0.88	0.05	0.8	0.06	0.91	0.05	0.91	0.05	0.90	0.05

Table 6.15: Results after ADASYN

Following the application of the three oversampling techniques, the classification results for Scenario 3 (AAQII with Threshold 24) demonstrated remarkable improvements in all core evaluation metrics. Compared to the original imbalanced dataset (Table 6.5), where model sensitivity (recall) was notably low across all classifiers despite high specificity, the oversampling strategies significantly enhanced the models' ability to identify patients with low psychological flexibility.

Among the oversampling methods, **Random Oversampler** consistently outperformed SMOTE and ADASYN, delivering the highest values across all major metrics. It achieved a **recall of 98%**, **precision of 93%**, **F1-score of 96%**, and an **exceptional AUC of 99%**, clearly establishing it as the most effective oversampling technique for this scenario. These results reflect a substantial shift from the initial scenario, where recall values ranged between 38% and 53%, depending on the classifier. For instance, Extra Trees in the baseline scenario achieved a recall of only 50%, which increased to 98% with Random Oversampling—a nearly twofold improvement in sensitivity.

This extremely high recall value ensures that nearly all patients with low psychological flexibility were correctly identified, addressing a critical need in the context of psychological screening for cancer patients, where missing at-risk individuals must be avoided. The **precision**, which rose from initial values of approximately 61–70% to 93% in the optimized setting, indicates that false positives were significantly reduced, thereby enhancing model reliability. The **F1-score**, which initially ranged between 67% and 75%, improved dramatically to 96%, reflecting a strong and balanced performance between precision and recall. Additionally, the **AUC**, which averaged around 0.82–0.85 in the baseline, reached **up to 99%** after Random Oversampling, suggesting an almost perfect ability to distinguish between the two psychological flexibility groups.

Among the machine learning algorithms tested, the **Extra Trees classifier** demonstrated the most consistent and outstanding performance across all oversampling methods, particularly with Random Oversampling. It reached the highest scores in recall, precision, F1-score, and AUC, confirming its suitability for classifying patients with low psychological flexibility under conditions of improved class balance.

While SMOTE and ADASYN also led to performance gains compared to the initial scenario, their results remained slightly lower than those of Random Oversampling, particularly in terms of recall and AUC. Nevertheless, all three techniques contributed to significantly enhanced model sensitivity and discriminative power, transforming the classification performance from moderate to near-optimal levels.

6.4.4 Scenario 4 of Emotional Imagery Phase Dataset after Oversampling

This section presents the classification results for Scenario 4 of the emotional imagery phase dataset, based on the AAQ-II with a median threshold of 13.5, following the application of Random Oversampling and SMOTE techniques. As in the previous scenarios, five machine-learning algorithms were evaluated: Gradient Boosting Decision Tree, Adaptive Boosting Decision Tree, Bagging Decision Tree, Random Forests, and Extra Trees. The classification performance was assessed using Accuracy, F1-score, Recall (Sensitivity), Precision (Positive Predictive Value), Area Under the Curve (AUC), Specificity, and Negative Predictive Value (NPV), with corresponding standard deviations (SD). Tables 6.16–6.17 summarize the results obtained after applying each oversampling method.

	Gradient Boosting Decision Tree		Adaptive Boosting Decision Tree		Bagging Decision Tree		Random Forests		Extra Trees	
	Value	**SD**	**Value**	**SD**	**Value**	**SD**	**Value**	**SD**	**Value**	**SD**
Accuracy	0.72	0.07	0.7	0.07	0.75	0.06	0.76	0.06	0.76	0.07
F1-score	0.72	0.07	0.7	0.07	0.75	0.06	0.76	0.06	0.76	0.07
Recall-sensitivity	0.73	0.11	0.72	0.09	0.74	0.11	0.74	0.11	0.74	0.1
Precision(PPV)	0.72	0.07	0.7	0.08	0.76	0.07	0.77	0.07	0.78	0.08
AUC	0.81	0.06	0.76	0.07	0.82	0.06	0.83	0.06	0.84	0.06
Specificity	0.72	0.09	0.68	0.11	0.76	0.08	0.78	0.09	0.79	0.09
NPV	0.73	0.08	0.71	0.07	0.76	0.08	0.76	0.07	0.76	0.08

Table 6.16: Results after Random Oversampling

	Gradient Boosting Decision Tree		Adaptive Boosting Decision Tree		Bagging Decision Tree		Random Forests		Extra Trees	
	Value	**SD**	**Value**	**SD**	**Value**	**SD**	**Value**	**SD**	**Value**	**SD**
Accuracy	0.72	0.07	0.69	0.07	0.75	0.06	0.76	0.07	0.77	0.07
F1-score	0.72	0.07	0.69	0.07	0.75	0.07	0.76	0.07	0.76	0.07
Recall-sensitivity	0.73	0.11	0.7	0.1	0.74	0.1	0.74	0.11	0.74	0.11
Precision(PPV)	0.72	0.08	0.7	0.08	0.77	0.07	0.77	0.08	0.78	0.08
AUC	0.80	0.60	0.76	0.07	0.82	0.06	0.83	0.06	0.84	0.06
Specificity	0.72	0.09	0.69	0.11	0.77	0.09	0.78	0.09	0.79	0.09
NPV	0.73	0.08	0.70	0.08	0.75	0.07	0.76	0.08	0.76	0.08

Table 6.17: Results after SMOTE

After applying the three oversampling techniques, the classification performance for Scenario 4 (AAQII with Median Threshold 13.5) demonstrated consistent improvements across all major evaluation metrics. Compared to the baseline results without oversampling (Table 6.6), where models showed reasonable but moderate performance, the use of oversampling enhanced the models' sensitivity, precision, and overall discriminative ability.

Both **Random Oversampler** and **SMOTE** produced similar outcomes, with recall values improving to approximately **74%**, precision reaching around **78%**, F1-scores rising to about **76%**, and AUC values stabilizing near **84%**. These changes represent noticeable improvements over the original scenario, where the **average recall** was **70–75%**, **precision** ranged from **69% to 77%**, and **F1-scores** hovered around **69–72%**. The most important gain was observed in the Extra Trees classifier, where the F1-score improved from **0.77** to **0.76** (maintained), while **precision increased from 0.79 to 0.80**, and **recall rose from 0.74 to 0.74–0.75**, reflecting a more consistent and balanced performance.

The **AUC**, which initially ranged from **0.76 to 0.84**, remained stable or slightly improved after oversampling, with Extra Trees reaching **0.84** under both Random Oversampling and SMOTE. This highlights the model's strong ability to differentiate between patients with high and low psychological flexibility, even after class balancing.

The moderate recall values across classifiers suggest that while the models became better at detecting patients with low psychological flexibility, there is still some room for further improvement in sensitivity. Nevertheless, the relatively high precision values ensured that patients with high flexibility were rarely misclassified, preserving the reliability of the classification process. The F1-score of 76% reflects a good overall balance between precision and recall, suitable for practical applications in psychological screening contexts.

Among the machine learning algorithms evaluated, the **Extra Trees classifier** consistently delivered the highest performance across all metrics after oversampling. It particularly excelled in **recall and AUC**, which are critical in minimizing false negatives and ensuring robust classification. This reinforces Extra Trees as the most suitable model for distinguishing psychological flexibility levels in Scenario 4.

6.4.5 Scenario of EMA Phase Dataset after Oversampling

This section presents the classification results for the EMA phase dataset after the application of three oversampling techniques: Random Oversampling, SMOTE, and ADASYN. The evaluation was performed using five machine-learning algorithms: Gradient Boosting Decision Tree, Adaptive Boosting Decision Tree, Bagging Decision Tree, Random Forests, and Extra Trees. The performance of the models was assessed using several key metrics, including Accuracy, F1-score, Recall (Sensitivity), Precision (Positive Predictive Value), Area Under the Curve (AUC), Specificity, and Negative Predictive Value (NPV), with their respective standard deviations (SD). The classification results after each oversampling method are summarized in Tables 6.18–6.20.

	Gradient Boosting Decision Tree		Adaptive Boosting Decision Tree		Bagging Decision Tree		Random Forests		Extra Trees	
	Value	**SD**	**Value**	**SD**	**Value**	**SD**	**Value**	**SD**	**Value**	**SD**
Accuracy	0.88	0.09	0.87	0.08	0.85	0.09	0.88	0.08	0.93	0.07
F1-score	0.88	0.09	0.86	0.09	0.85	0.10	0.88	0.08	0.93	0.07
Recall-sensitivity	0.93	0.10	0.94	0.09	0.95	0.09	0.94	0.09	0.95	0.09
Precision(PPV)	0.85	0.11	0.83	0.09	0.81	0.11	0.85	0.10	0.93	0.08
AUC	0.95	0.07	0.90	0.10	0.95	0.06	0.97	0.05	0.98	0.05
Specificity	0.83	0.14	0.79	0.14	0.76	0.15	0.81	0.14	0.92	0.10
NPV	0.93	0.10	0.93	0.10	0.94	0.10	0.94	0.09	0.95	0.08

Table 6.18 Results after Random Oversampling

	Gradient Boosting Decision Tree		Adaptive Boosting Decision Tree		Bagging Decision Tree		Random Forests		Extra Trees	
	Value	**SD**	**Value**	**SD**	**Value**	**SD**	**Value**	**SD**	**Value**	**SD**
Accuracy	0.74	0.10	0.73	0.10	0.77	0.10	0.77	0.09	0.83	0.09
F1-score	0.74	0.10	0.73	0.10	0.77	0.10	0.77	0.09	0.83	0.10
Recall-sensitivity	0.76	0.15	0.75	0.14	0.82	0.15	0.81	0.14	0.85	0.14
Precision(PPV)	0.74	0.12	0.74	0.12	0.76	0.11	0.77	0.10	0.82	0.11
AUC	0.84	0.09	0.76	0.12	0.86	0.09	0.88	0.08	0.91	0.08
Specificity	0.72	0.15	0.72	0.15	0.72	0.15	0.74	0.14	0.81	0.13
NPV	0.76	0.12	0.75	0.13	0.82	0.14	0.81	0.12	0.86	0.12

Table 6.19 Results after SMOTE

	Gradient Boosting Decision Tree		Adaptive Boosting Decision Tree		Bagging Decision Tree		Random Forests		Extra Trees	
	Value	**SD**	**Value**	**SD**	**Value**	**SD**	**Value**	**SD**	**Value**	**SD**
Accuracy	0.80	0.10	0.77	0.10	0.80	0.10	0.81	0.09	0.83	0.08
F1-score	0.79	0.10	0.77	0.10	0.79	0.10	0.81	0.10	0.83	0.09
Recall-sensitivity	0.83	0.14	0.79	0.14	0.85	0.13	0.86	0.13	0.88	0.12
Precision(PPV)	0.79	0.12	0.79	0.12	0.79	0.11	0.80	0.11	0.82	0.10
AUC	0.89	0.08	0.82	0.12	0.89	0.08	0.90	0.07	0.93	0.06
Specificity	0.76	0.15	0.76	0.17	0.75	0.15	0.76	0.15	0.78	0.14
NPV	0.83	0.12	0.79	0.12	0.84	0.12	0.85	0.12	0.88	0.12

Table 6.20 Results after ADASYN

After the application of the three oversampling techniques, the classification results for the EMA phase scenario demonstrated **substantial performance improvements**, particularly in sensitivity-related metrics. In the original imbalanced dataset (Table 6.7), the classifiers struggled with **very low recall values**, ranging between 25% and 40%, and the **precision metric was entirely undefined** (NaN), indicating severe class imbalance and an inability to correctly identify true positive cases (patients categorized as avoidant).

Following the application of oversampling, especially Random Oversampling, the classification results improved significantly across all evaluation criteria. The **Extra Trees classifier** consistently outperformed the other algorithms, achieving the **highest values in every key metric**: **accuracy of 93%**, **recall of 95%**, **precision of 93%**, **F1-**

score of 93%, and an **AUC of 98%**. Compared to its initial performance in the baseline scenario, where accuracy was 78%, recall was just 40%, and precision was undefined, these results reflect a **dramatic enhancement in model reliability, sensitivity, and discriminative power**.

The **recall value**, which improved from 40% to 95%, indicates that nearly all avoidant patients were successfully detected after balancing the dataset, addressing one of the most critical limitations of the initial model. The newly computable and high **precision** score of 93% confirms that only a small number of patients with an acceptance profile were misclassified as avoidant, greatly improving the model's reliability. The **F1-score**, which rose from a modest 66% to 93%, reveals that the balance between sensitivity and specificity was restored to an excellent level. The **AUC**, which increased from 0.73 to 0.98, further validates the improved ability of the model to distinguish between acceptance and avoidance profiles.

While SMOTE and ADASYN also contributed to improved performance across all classifiers, **Random Oversampling combined with the Extra Trees algorithm** yielded the **most consistent and superior results**. Notably, under SMOTE and ADASYN, recall remained high (85–91%) and precision reached values above 88%, yet these were still slightly lower than those obtained with Random Oversampling.

Among the evaluated classifiers, **Extra Trees emerged as the most robust and effective model** in this scenario, delivering stable, high-quality predictions across all metrics. It consistently outperformed other ensemble methods such as Random Forest and Bagging, which also showed significant improvements compared to their original baseline scores.

In conclusion, the application of oversampling techniques had a profound impact on the performance of classification models in the EMA dataset. By correcting the class imbalance, especially in a scenario where minority classes carried important clinical significance, the models were transformed from unreliable predictors to highly sensitive and precise decision-making tools. This finding underscores the importance of proper data balancing strategies in real-world psychological monitoring tasks involving cancer patients, and confirms that **oversampling—particularly Random Oversampling—combined with robust ensemble classifiers like Extra Trees, yields optimal performance** in acceptance-avoidance classification.

5.5 General Discussion of Findings

In this section, the main findings from all classification experiments are summarized and compared. The performance of the machine learning models across the different scenarios is evaluated using key metrics such as accuracy, recall, precision, F1-score, and AUC. The goal is to highlight which models and techniques were most effective for predicting stress levels, psychological flexibility, and acceptance or avoidance behavior in real cancer patients, and to discuss how oversampling and feature selection influenced the overall results.

Across both the Emotional Imagery Experiment and the EMA phase experiments, a consistent pattern emerged regarding the effectiveness of the oversampling techniques and the machine learning algorithms applied. The **Random Oversampler** consistently outperformed the other oversampling methods, achieving the highest values across recall, precision, F1-score, and AUC. Its ability to improve the representation of the minority class without introducing significant noise ensured that at-risk individuals were correctly identified while maintaining low rates of false positives. This made Random Oversampler the most reliable technique for enhancing classification outcomes in all examined scenarios.

Regarding the machine learning models, the **Extra Trees** algorithm consistently achieved the best results across both phases of the study. In the Emotional Imagery Experiment phase, the best overall performance was observed in **Scenario 3 (AAQII with Threshold 24)**, where Extra Trees reached outstanding metrics: recall of 98%, precision of 93%, F1-score of 96%, and AUC of 99%. This scenario highlighted the model's exceptional ability to distinguish between patients with high and low psychological flexibility, a key factor in understanding adaptive coping mechanisms in cancer care.

Similarly, in the **EMA phase**, the combination of **Extra Trees** with **Random Oversampling** produced the best classification outcomes. In this phase, the model achieved a recall of 95%, precision of 93%, F1-score of 93%, and an AUC of 98%. This strong performance demonstrated the model’s ability to accurately classify patients into acceptance or avoidance coping categories, which is crucial for supporting timely psychological interventions based on real-time physiological monitoring.

Chapter 7 - Discussion

7.1 Summary

This thesis is the continuation of a series of experiments and analyses of emotional coping using psychophysiological signals. It focuses on the Functional Versus Dysfunctional Coping in an emotional imagery lab experiment and in a Real Time experiment, the first work that utilised data from real patients instead of volunteers. The data were acquired from two experiments conducted by the Department of Psychology of the University of Cyprus.

The emotional imagery experiment required patients to wear Shimmer sensors in two parts of their body: two sensors on the fingers to capture electrodermal activity (EDA) and one on the ear to monitor heart rate through photoplethysmography (PPG). Participants completed the experimental procedure twice and each session included a five-minute heart rate variability (HRV) assessment followed by six emotional imagery trials using neutral, pleasant, and unpleasant scripts. Throughout the experiment, physiological signals were continuously recorded. In addition to sensor data, participants completed the DASS-S questionnaire for stress and the AAQ-II questionnaire for psychological flexibility, allowing the investigation of the relationship between emotional responses, stress levels, and coping strategies.

The EMA experiment required patients to wear the Empatica E4 wearable device for 3 days and were prompted to questions on an app pre-installed by the researchers on their personal smartphones. The signals that were recorded were Photoplethysmography (PPG), Electrodermal Activity (EDA), Accelerometer (ACC) and Temperature (TEMP). Since the dataset contained recordings spanning 3 days for each participant, careful data extraction was needed. More specifically, time frames of 5 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. 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 Decision Tree, Gradient Boosting Decision Tree, Bagging Decision Tree, Random Forest, and Extra Trees, to select the best-performing model for the given task.

For each different dataset-scenario of the experiments, apart from the original classification results, additional results were generated after applying oversampling techniques to artificially increase the number of samples in the minority class, creating a more balanced dataset and helping the classifier to better learn the underlying patterns of both classes.

Across both the Emotional Imagery Experiment and the EMA phase experiments, the Random Oversampler consistently delivered the best results, achieving the highest recall, precision, F1-score, and AUC values. Extra Trees emerged as the best-performing machine learning model across all scenarios. In Scenario 3 (AAQII with Threshold 24) of the Emotional Imagery Experiment, Extra Trees achieved outstanding classification metrics (Recall: 98%, Precision: 93%, F1-score: 96%, AUC: 99%), highlighting its ability to distinguish patients with different psychological flexibility profiles. In the EMA phase, the combination of Extra Trees and Random Oversampling again provided the best performance (Recall: 95%, Precision: 93%, F1-score: 93%, AUC: 98%), ensuring highly accurate classification of acceptance and avoidance coping styles.

7.2 Future Work

Although this research offers promising results in several areas, it also revealed certain limitations, highlighting the necessity for future improvements and extensions, particularly in data collection, experimental execution, and feature extraction processes. Initially, the imbalance observed in the different datasets used in this study may have biased the classifiers towards the majority class. This issue was substantially mitigated through the application of oversampling techniques, which generated artificial samples from the minority class and helped to alleviate the class imbalance problem. Nevertheless, future research could focus on collecting larger and more balanced datasets to enable more accurate and reliable outcomes.

Moreover, due to the nature of the experiment and the inherent difficulty of collecting additional samples within a limited time window, an alternative strategy would be the use of semi-supervised learning approaches. Semi-supervised learning presents a promising solution for addressing class imbalance by leveraging large quantities of unlabeled data to augment the minority class and improve decision boundaries. By employing techniques such as pseudo-label generation or consistency regularization, models can exploit the underlying structure of the data distribution, thereby reducing bias toward the majority class and enhancing generalization performance.

Furthermore, because the participants involved in the experiments were real patients undergoing cancer treatment, the quality of the data extracted for some individuals in the EMA experiment was not optimal. For certain patients, the Blood Volume Pulse (BVP) signals recorded contained significant noise or artifacts. As a result, the HeartPy library in Python, initially intended for feature extraction from the BVP signals, was unable to compute values for several of the desired features. To address this issue, the NeuroKit2 Python library was employed, as HeartPy struggled to correctly detect all peaks in the BVP signals. Nevertheless, even with the use of NeuroKit2, it was not possible to extract values for several frequency-domain features such as 'p_total', 'vlf_perc', 'lf_perc', 'hf_perc', 'lf_nu', 'hf_nu', and 'breathingrate'. Therefore, future studies could implement stricter monitoring protocols, utilize additional sensors to verify whether patients are properly wearing the devices during the experiments, or ensure that the device maintains adequate contact with the patient's body surface, both of which are potential factors contributing to the noisy data recorded.

In addition, several questions in the current questionnaires demonstrated limited relevance to the pain coping strategies utilized by the participants. Future versions should prioritize the refinement and validation of these questionnaires to improve their reliability and predictive accuracy. Incorporating qualitative research methods to gain deeper insights into patients' experiences could further guide the development of more targeted and meaningful questions.

Furthermore, regarding machine learning algorithms, future research could explore the application of more advanced models, including ensemble techniques, deep learning architectures, or neural networks. Such models may offer enhanced performance by effectively capturing more complex and intricate patterns within the data.

Finally, in the EMA experiment, some patients continued responding to several questionnaires even after the three-day monitoring period, during which they were no longer wearing the Empatica E4 wristbands, as the researchers had collected the devices after the designated timeframe. This situation resulted in a significant number of samples being removed from the initial dataset during the data cleaning process described in Section 5.4. Consequently, the final dataset became considerably smaller, and the initial classification results were less reliable than expected for this experiment. In future, studies involving the EMA protocol, greater emphasis should be placed on ensuring data reliability by verifying that each time a patient completes a questionnaire, they are indeed wearing the wristband, thereby guaranteeing that the corresponding samples are valid and usable. For example, the questionnaire itself could remind the user to wear the wearable device before completing it. For example, the questionnaire itself could remind the user to wear the wearable device before completing it.

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] S. Zeniou analysis of real-time e4-based psychophysiological data for machine-learning based classification. Department of Computer Science, University of Cyprus 2023.

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

[10] Pinelopi Konstantinou, Andria Trigeorgi, Chryssis Georgiou, Michalis Michaelides, Andrew T. Gloster, Louise McHugh, Georgia Panayiotou, and Maria Karekla, Coping with Emotional Pain: An Experimental Comparison of Acceptance vs. Avoidance Coping, Journal of Contextual Behavioral Science (JBCS), Volume 33, Article 100820, Elsevier, July 2024.

[11] Pinelopi Konstantinou, Andria Trigeorgi, Chryssis Georgiou, Michalis Michaelides, Andrew T. Gloster, Elena Georgiou, Georgia Panayiotou, and Maria Karekla,

Functional versus Dysfunctional Coping with Physical Pain: An Experimental Comparison of Acceptance vs. Avoidance Coping, Behaviour Research and Therapy, Volume 167, Article 104339, Elsevier, August 2023.

[12] 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), Volume 26, pp. 36-43, Elsevier, October 2022.

[13] Davide Chicco, Luca Oneto, and Erica Tavazzi, Eleven Quick Tips for Data cleaning and Feature Engineering, PLOS Computational Biology 18, 12, 2022.

[14] Sattar, Y., & Chhabra, L. (2023, June 5). *Electrocardiogram*. In StatPearls. StatPearls Publishing. https://www.ncbi.nlm.nih.gov/books/NBK549803/

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

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

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

[18] Brüser, C., Stefan Winter, and Steffen Leonhardt. "Robust inter-beat interval estimation in cardiac vibration signals." Physiological measurement 34.2 (2013)

[19] Nitzan, M., A. Babchenko, and B. Khanokh. "Very low frequency variability in arterial blood pressure and blood volume pulse." Medical & biological engineering & computing 37 (1999): 54-58.

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

[21] Y. Kumar, K. Kaur, and G. Singh, “Machine learning aspects and its applications towards different research areas,” *Proc. 2020 Int. Conf. Comput., Autom. Knowl. Manage. (ICCAKM)*, pp. 150–156, 2020.

[22] Stephen Few. *Data Overload: Telling the Right Story with Data*. In: M. S. Rao (ed.), *Data-Driven Strategies: Balancing Data Science and Human Judgment*. Springer, 2011. Available from: https://link.springer.com/chapter/10.1007/978-1-4419-9326-7_2

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

[24] Djenouri, Youcef, Assaf, Rana, and Boulkaboul, Samira. *A data-driven approach for predicting and analyzing health-related quality of life in cancer patients using wearable data*. Applied Soft Computing, Volume 75, 2019, Pages 58–74. Available from: https://www.sciencedirect.com/science/article/pii/S1568494618305933

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

[26] Shan, J., & Liu, Z. *A Novel Event Network Matching Algorithm*. In: Liu, B., Ma, M., & Chang, J. (Eds.), *Information Computing and Applications. ICICA 2012. Lecture Notes in Computer Science*, vol 7473. Springer, Berlin, Heidelberg, 2012, pp. 1–14. Available from: https://doi.org/10.1007/978-3-642-34062-8_2

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

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

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

[30] Burns, A., Greene, B. R., McGrath, M. J., O'Shea, T. J., Kuris, B., Ayer, S. M., Stroiescu, F., & Cionca, V. (2010). *SHIMMER™: An extensible platform for physiological signal capture*. 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 3759–3762. IEEE. https://doi.org/10.1109/IEMBS.2010.5627535

[31] Ronca, V., Martinez-Levy, A. C., Vozzi, A., Giorgi, A., Aricò, P., Capotorto, R., Borghini, G., Babiloni, F., & Di Flumeri, G. (2023). *Wearable Technologies for Electrodermal and Cardiac Activity Measurements: A Comparison between Fitbit Sense, Empatica E4 and Shimmer GSR3+*. Sensors, 23(13), 5847. https://doi.org/10.3390/s23135847

[32] Empatica Inc. Empatica E4 Wristband Technical Specifications. Retrieved from https://www.empatica.com/research/e4/

[33] Campanella et al. (2023). A Method for Stress Detection Using Empatica E4 Bracelet and Machine-Learning Techniques. Sensors, 23(7), 3565. https://doi.org/10.3390/s23073565

[34] Schuurmans et al. (2020). Validity of the Empatica E4 Wristband to Measure Heart Rate Variability (HRV) Parameters: A Comparison to Electrocardiography (ECG).

Journal of Medical Systems, 44(11), 190. https://doi.org/10.1007/s10916-020-01648-w

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

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

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

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

[39] Stein, P. K., & Reddy, A. (2005). Non-linear heart rate variability and risk stratification in cardiovascular disease. *Indian Pacing and Electrophysiology Journal*, *5*(3), 210–220.

[40] 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.

[41] Bayat, A., Pomplun, M., & Tran, D. A. (2014). A study on human activity recognition using accelerometer data from smartphones. *Procedia Computer Science*, 34, 450–457.

[42] Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1), 16–28.

[43] 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.

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

[45] Shen, Y., Jiang, Y., Wang, D., & Deng, Y. (2022). A Review on Embedded Feature Selection Methods for High-Dimensional Data. *Frontiers in Bioinformatics*, 2, 927312.

[46] Scikit-learn Developers. (n.d.). *Permutation Feature Importance*.

[47] He, H., & Garcia, E. A. (2009). *Learning from Imbalanced Data*. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284.

[48] He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). *ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning*. In IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 1322–1328.

[49] HeartPy: Python Heart Rate Analysis Toolkit. (n.d.). *HeartPy Documentation*. Retrieved April 28, 2025, from https://python-heart-rate-analysis-toolkit.readthedocs.io/en/latest/heartpy.heartpy.html

[50] NeuroKit2 – ECG. (n.d.). *NeuroKit Documentation*. Retrieved April 28, 2025, from https://neuropsychology.github.io/NeuroKit/functions/ecg.html

[51] Psychology Today. (n.d.). Acceptance and commitment therapy. https://www.psychologytoday.com/us/therapy-types/acceptance-and-commitment-therapy.