Individual Diploma Thesis

FEATURE SELECTION AND TRAINING OF MACHINE LEARNING ALGORITHMS TO CLASSIFY FUNCTIONAL VERSUS DYSFUNCTIONAL COPING WITH ACUTE PAIN

Eleni Georgiou

UNIVERSITY OF CYPRUS

DEPARTMENT OF COMPUTER SCIENCE

May 2022

UNIVERSITY OF CYPRUS

DEPARTMENT OF COMPUTER SCIENCE

Feature Selection and Training of Machine Learning Algorithms to Classify Functional versus Dysfunctional Coping with Acute Pain

Eleni Georgiou

Supervisor

Chryssis Georgiou

The Individual Diploma Thesis was submitted for partial fulfilment of the requirements for the degree of Computer Science of the Department of Computer Science of the University of Cyprus

May 2022

Acknowledgements

Foremost, I would like to express my sincere gratitude to my supervisor, Dr Chryssis Georgiou, Professor at the Department of Computer Science at the University of Cyprus, for trusting me on this topic and for the support throughout my entire journey toward my Diploma Thesis.

I would also like to thank the Department of Psychology of the University of Cyprus, and especially Dr Maria Karekla for providing us with the experimental data samples that we used.

Furthermore, my sincere thanks go to Andria Trigeorgis who provided me with important background knowledge based on previous work on the topic and gave me continuous guidance and feedback throughout this project.

I am deeply grateful to my parents and friends for their encouragement. Without them, this study would not be possible. Finally, I would like to thank my dog, Brad, for being so loyal, loving, and always by my side.

Abstract

During the last years, there has been a rapid development of both wearable technologies and Artificial Intelligence. This thesis aims to bring these two together to prevent dysfunctional pain-coping attempts. More precisely, this thesis aims to examine whether wearable devices are able to correctly classify individuals into functional or dysfunctional regarding pain coping.

In an experiment conducted by the Department of Psychology of the University of Cyprus, all participants were asked to immerse their hands into an ice water container. During the experiment, a variety of psychophysiological signals — electrocardiogram (ECG), electrodermal activity (EDA), and facial electromyography (fEMG) — were recorded from the participants, both from BIOPAC, which is a stationary device and from two wearable devices: Microsoft Band 2 and Moodmetric Smart Ring. Having the raw signals available, the Rectangular Window Methodology (RWM) was utilized in order to increase the size of the dataset. Thereafter, multiple psychophysiological features were extracted. Using three different feature selection methods (Exhaustive Feature Selection, Feature Importance and Correlation Coefficient), the most significant features were chosen. Using the selected features, a number of Machine Learning models (Adaptive Boosting, Gradient Boosting Decision Tree, Random Forest, and Extra Trees) were trained and compared in order to find the most effective one.

The analysis shows that the most important features, in order to effectively classify people into the two categories, are average heart rate and peak amplitude of Skin Conductance Response (SCR). The Gradient Boosting Decision tree, when fine-tuned and combined with RWM and a window of 10 seconds, can correctly classify people, with accuracy reaching 85%, making it a very powerful model.

In a comparison regarding the monitoring devices, it was shown that the findings are different when using different methods to compare the data. Pearson correlation shows no correlation between most of the features, while Bland-Altman scatter plots show agreement in the measurements in most of the cases. Furthermore, by comparing the results of the current study with the previous studies it was observed that when using RWM the accuracy of the classifiers

increases. Moreover, data acquired by the band can give results with performance similar to those acquired by the stationary device.

Table of Contents

Chapter 1 Introduction 1

1.1 Motivation 1

1.2 Goals of the Study 2

1.3 Methodology 2

1.4 Document Organization 4

Chapter 2 Background Knowledge 6

2.1 Psychophysiological Signals 6

2.2 Machine Learning Algorithms 8

2.3 Model Evaluation 11

2.4 Monitoring Devices 14

Chapter 3 Data Collection and Previous Work 16

3.1 Physiological Experiments 17

3.2 Related Work 19

Chapter 4 Signal Analysis 22

4.1 Training Data Multiplication 22

4.2 Feature Extraction 24

4.3 Feature Selection 30

Chapter **5** **Classification 45**

5.1 Classifier Fine-Tuning 45

5.2 Classifier Selection 47

Chapter **6** **Comparisons 49**

6.1 Comparison of Psychophysiological Features Between Devices 49

6.2 Comparison of Results with Previous Work 56

Chapter **7** **Discussion** **59**

8.1 Summary 59

8.2 Future Work 57

References **62**

Chapter 1

Introduction

1.1 Motivation 1

1.2 Goals of the Study 2

1.3 Methodology 2

1.4 Document Organization 4

1.1 Motivation

The use of wearable devices, such as smartwatches and smart bands, has greatly increased, over the past decade. This kind of devices is able to record multiple measures, including heart and sweat gland activity. These measures are called psychophysiological signals and are proven to reflect an individual's emotional arousal. Examples of such psychophysiological signals are Electrocardiogram (ECG), Electrodermal Activity (EDA), and facial electromyography (fEMG).

A number of previous studies [1] [2] [3] analysed such data. Due to the fact that this topic is wide and multiple aspects need to be investigated, there are some areas that need further research. Firstly, previous work concentrated on signals recorded from stationary devices, while wearable devices' data were not examined. Moreover, the only features that were examined and used to train the models, are HRV time-domain features that come from ECG. Last but not least, not enough focus was given in feature selection techniques, which is a really important process in order to obtain good results. This study aims to investigate further this topic, in order to assist health care.

1.2 Goals of the Study

This study makes use of data collected from an experiment regarding pain management techniques that was conducted by the Department of Psychology of the University of Cyprus. The ultimate goal of the present thesis is to contribute to the integration of a form of psychotherapy, called Acceptance and Commitment Therapy, in the everyday life. *Acceptance and Commitment Therapy* (ACT) [4] involves encouraging people to try to deal with their thoughts and feelings instead of blaming themselves about them or trying to ignore them. ACT is a vital therapy as it can help people struggling with OCD, anxiety, depression, etc. It separates people into two groups, based on their reactions at a certain time. The first group is 'acceptance', also known as 'functional', and it contains people who accept their problems and try to cope with them head-on. Contrary, the 'avoidance' group, which is also known as 'dysfunctional', is the exact opposite and includes people that deny to remain in contact with their thoughts and sensations and attempt to avoid them. An individual does not always fall into the same category; their classification changes depending on the environment and the circumstances. This thesis aims to effectively classify individuals into functional or dysfunctional regarding pain coping.

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

1.3 Methodology

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

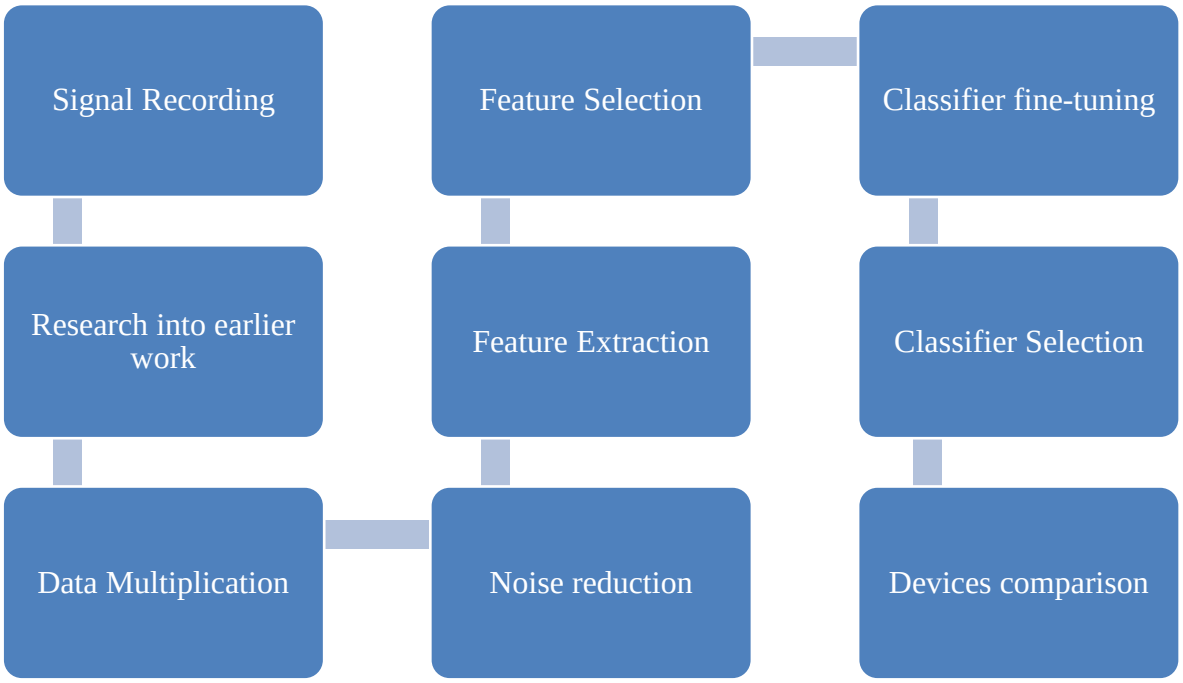


Figure 1.1 Methodology of current work

At first, the Department of Psychology of the University of Cyprus conducted an experiment in the lab, where psychophysiological **signals were recorded** both from stationary and wearable devices. These signals are electrocardiogram (ECG), Heart Rate Variability (HRV), facial electromyography (fEMG), and electrodermal activity (EDA).

Prior to analysing the collected data, special emphasis was placed on **understanding the work done** during the previous years. Thus, all methodologies employed were considered and the most effective ones were applied in the present work.

Due to the fact, that the data samples collected were insufficient, the Rectangular Window Methodology [5] was used, in order to **produce more data samples**. This was done in windows of four different sizes: 10, 20, 30, and 40 seconds.

Thereafter, using Python, each signal was **cleaned from noise** and artefacts caused by device errors. The raw signals cannot be used directly. Thus, the next step was to **extract features** from each one of the psychophysiological signals. The focus was given to time-domain and statistical features.

In order to train the algorithms in a reasonable amount of time, only the most important features needed to be selected. To do this, techniques from three different **Feature Selection** methods were applied and compared. The techniques examined are Exhaustive Feature Selection, Feature Importance and Correlation Coefficient.

Five different supervised binary classification Machine Learning algorithms were examined. These are Adaptive Boosting, Gradient Boosting Decision Tree, Bagging Decision Tree, Random Forest, and Extra Trees. These algorithms have some parameters that need to be tuned, depending on the data, in order to perform well. Thus, **classifier fine-tuning** followed, where different values of the parameters of each algorithm were tested, and the best one was chosen. The next step was to **train the five algorithms**, using the parameters chosen in the previous step, and select the one with the best performance.

Finally, some comparisons were performed. Firstly, the **signals** recorded from the three devices (stationary, band, ring) **were compared** in order to find out whether wearable devices are reliable to record psychophysiological signals. Then, the results of the current work were compared to the results of previous related works.

1.4 Document Organization

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

Chapter Number	Chapter Description
2	Overviews the background knowledge on which the thesis was built on. At first, the psychophysiological signals that were recorded in the Psychology lab are explained. Thereafter, the Machine Learning algorithms used are analysed., as well as the methodology and metrics used for evaluating the models. Finally, the devices that were used throughout the experiments are briefly described.
3	Explains in detail the three previous experiments.
4	Main chapter. Explains all the work that was done to obtain the features used to train the algorithm from the raw signals. Firstly, it

Chapter Number	Chapter Description
	explains the methodology used to increase the number of data samples. Afterwards, it clarifies the structure of each signal and the features that were extracted from them. At last, it analyses three important feature selection methods and concludes to the selected subset of features.
5	Describes how the classifiers were tuned, meaning how the values of their parameters were picked, in order to perform best. Also, it presents the performance of each algorithm and concludes to the most effective one.
6	Describes the comparisons that were performed. Compares the signals recorded from the three different devices, BIOPAC, Microsoft Band 2, and Moodmetric Smart Ring. Moreover, this chapter compares the results of the current work with the results of previous related works.
7	Summarizes the work done and suggests future improvements.

Table 1.1 Document Organization

Chapter 2

Background Knowledge

2.1 Psychophysiological Signals 6

2.2 Machine Learning Algorithms 8

2.3 Model Evaluation 11

2.4 Monitoring Devices 14

2.1 Psychophysiological Signals

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

2.1.1 Electrocardiogram (ECG)

When the heart beats, it produces electrical signals. These electrical signals can be recorded non-invasively from the body surface using an *electrocardiogram* (ECG) [6]. The basic pattern of this electrical activity comprises three waves, which have been named P, QRS, and T [7]. A more detailed explanation of how the ECG signal is formed, and what features can be extracted from it, can be found in Section 4.2.1

2.1.2 Photoplethysmography (PPG)

During a cardiac cycle, i.e., from the beginning of a heartbeat to the beginning of the next one, blood volume rises and falls throughout the body. The difference in blood volume can be observed in the skin's outer layers and be measured using optical sensors [8]. More specifically, *photoplethysmography* (PPG) is a technology that uses an LED light source and a photodetector. LED emits light into the microvascular bed of tissue and the photodetector, which is a light-sensitive sensor, records how much light is absorbed or reflected. The amount of light that is absorbed or reflected changes based on blood volume [8] [9]. From photoplethysmography signal Heart Rate Variability (HRV) can be estimated, which is equal to the distance between consecutive R-peaks of the ECG signal. Section 4.2.1 [10] explains in detail the features that can be extracted from HRV.

2.1.3 Electrodermal Activity (EDA)

Electrodermal activity (abbreviated as EDA), also known as *Galvanic Skin Response* (GSR) or *Skin Conductance* (SC), is the change of the electrical properties of the skin which follows sweating. It reflects the intensity of a person's emotional state or emotional arousal. The variation of skin conductance can be measured non-invasively by applying an electrical potential between two points on the skin and measuring the current flow between them. EDA is linked to emotional arousal and clinical applications cover a wide range of topics, including pain evaluation [11] [12]. A more detailed explanation of how the EDA signal is formed, and what features can be extracted from it, can be found in Section 4.2.2.

2.1.4 Facial Electromyography (fEMG)

Facial Electromyography (abbreviated as fEMG) is a technique that is used to detect emotional expressions by recording the movement of the muscles on the face. Each time a muscle contracts, a burst of electric activity is produced and propagated through adjacent tissue and bone, which can be recorded from neighbouring skin areas. The muscles whose activity is most commonly recorded are zygomaticus major (ZYG), corrugator supercilii (COR) and orbicularis oculi (ORB) [13].

Zygomaticus major is placed on the region of the cheek, is responsible for smiling and is associated with positive emotional stimuli. Corrugator supercilii is placed above the brow, as it is involved in brow furrowing and is linked with negative reactions [13] [14]. The orbicularis oculi muscle is found on the eyelids and its primary purpose is to close the eyelids [15]. According to a study, even when participants are instructed to hide their facial expressions, fEMG is still able to detect muscle activation [16].

2.2 Machine Learning Algorithms

There are three types of Machine Learning algorithms: *supervised*, *unsupervised*, and *reinforcement* [17]. In supervised algorithms, each sample in the dataset has a label, which corresponds to the target, and the algorithm attempts to find patterns with the ultimate goal of being able to generalize between cases that do not belong to the dataset. Such methods are usually called "learning with a teacher". In unsupervised techniques, the desired output for each input is not known, so such algorithms attempt to find common features between the input data. Finally, reinforcement learning is similar to “training with a judge”, since the algorithm gets a reward every time it returns a correct result, while in the opposite case it gets a penalty.

The algorithms that will be examined in this thesis belong to the category of supervised learning, since for each sample of the dataset, the desired label was assigned by a human (in our case a psychologist). More specifically, algorithms based in decision trees will be used, as previous work has shown that they perform well in similar problems [18] [3]. A decision tree is an effective classification tool. Its structure is similar to a tree. Each internal node corresponds to a test (i.e., decision) on an input variable, while each edge that comes out of a certain node indicates an outcome of the test. Leaf nodes hold a class label [19].

This subsection explains the Machine Learning algorithms that will be trained for classification.

2.2.1 AdaBoost Algorithm

AdaBoost Algorithm [20], which is also known as Adaptive Boosting, uses a sequence of weak learners. A weak learner is a model that performs at least slightly better than random guessing. By combining a set of weak learners, a strong learner can be achieved.

This algorithm assigns a weight to each sample of the dataset. Initially, the weight of each sample is equal to $\frac{1}{N}$, where N is the total number of samples. Therefore, all samples are considered equally important. In the next iteration, the weights of each sample change, depending on whether the sample was correctly classified or not. In case it was not correctly classified, then its weight increases, while in the opposite case it decreases. Thus, in each boosting iteration, the algorithm is forced to focus on data samples that were wrongly classified. This process is repeated until the desired accuracy is achieved or until a predefined number of learners are used. The learner commonly used is a Decision Tree with only one level, meaning that it only makes one decision.

2.2.2 Gradient Boosting Decision Tree

Gradient Boosting Decision Trees (GBDT) [21] has three basic components. These are a weak learner, a loss function and an additive model. In each iteration of the algorithm, a new weak learner (i.e., a decision tree) is added to the model. Afterwards, the loss function is computed, which makes an estimation of how efficient the model is based on the expected and the real output values of the classifier. In the next iteration, a new decision tree is added to the model in order to decrease the value of the loss function.

2.2.3 Bagging Decision Tree

Bagging Decision Tree is based on the *Bagging Ensemble technique* [22] (Figure 2.1), which combines a set of decision trees to come up with a result.

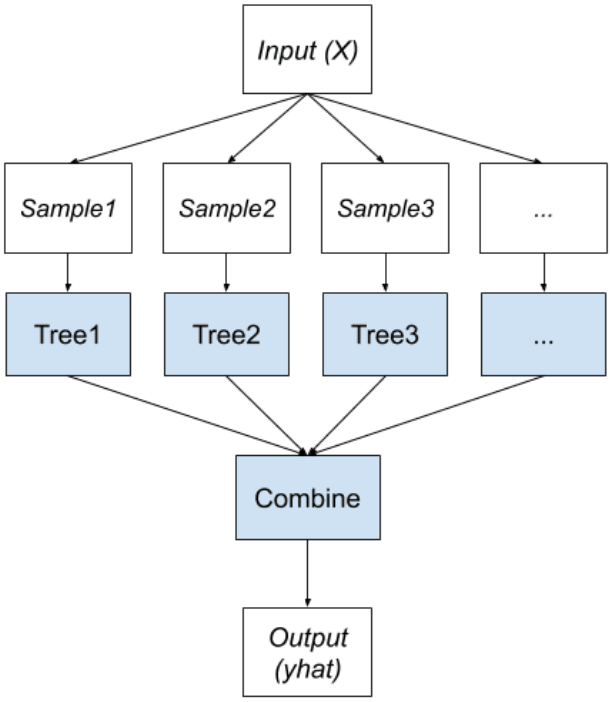


Figure 2.1 Bagging Ensemble Technique [22]

At first, a collection of decision trees is created using all the features available in the dataset and is trained using a random subset of the training data. When a new sample arrives, each of the decision trees is used to classify it into a class. The class returned by the algorithm is the one that was voted by the majority of the classifiers.

2.2.4 Random Forest

Random Forest (RF) algorithm [23] belongs to the category of Bagging Ensemble algorithms. Each tree is created by selecting a number of random samples from the dataset with replacement and then selecting a set of attributes (without replacement). This procedure is repeated until a forest consisting of a predefined number of trees is created. The final class labels are based on majority voting of the decision trees in the forest. By increasing the number of trees in the forest, the performance of the algorithm improves, whereas more computational power is needed. An important property of Random Forest algorithm is that it does not overfit when more trees are added.

2.2.5 Extra Trees

Extra Trees algorithm [24] is very similar to Random Forest, with the difference that Extra Trees add randomization. While Random Forest chooses the optimum split in a node (i.e., the optimum range of the node's feature value for each branch), this algorithm chooses it randomly. When the splits are performed for all candidate features, those that had the best performance are chosen.

2.3 Model Evaluation

To determine the most-effective classification algorithm in the topic of this thesis, it is important to select the most suitable evaluation methodology and performance metrics against which the candidate algorithms will be compared.

2.3.1 Evaluation Methodology

The evaluation methodology used to compare the performance of the Machine Learning algorithms is *Stratified k-fold cross-validation* [25], which was used in related works in the past [3] [1] [2]. This method has a single parameter that needs to be chosen, and that is *k*. The data are firstly split into *k* different class-balanced groups. By class balanced it is meant that the proportion of the samples from the two classes is the same. Following, there are *k* iterations. In each iteration, one of the groups is used as the test set and the rest *k*-1 groups are used as the training set to fit a model. The model is then evaluated using the test set and the model is discarded. At the end of the *k* iterations, the average evaluation score is returned.

2.3.2 Performance Metrics

To compute the performance metrics [26], four important measures are needed. These are true positives, false positives, true negatives and false negatives, where:

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

- False Negatives (FN): The number of samples incorrectly classified as negative

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

Confusion Matrix

The confusion matrix is a square matrix that includes the four measures explained above. The format of a confusion matrix is shown in Figure 2.2.

		Predicted	
		Negative (**N**) -	Positive (**P**) +
Actual	Negative -	True Negatives (**TN**)	False Positives (**FP**) **Type I error**
	Positive +	False Negatives (**FN**) **Type II error**	True Positives (**TP**)

Figure 2.2 Confusion matrix format

Accuracy

Correct predictions to total predictions ratio.

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

PPV (Precision)

True positives to total predicted positives ratio.

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

NPV

True negatives to total predicted negatives ratio.

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

Specificity

True negatives to total negatives in the data ratio.

$$NPV = \frac{TN}{TN + FP}$$

Sensitivity (Recall)

True positives to total positives in the data ratio.

$$Sensitivity = \frac{TP}{TP + FN}$$

F1-score

The harmonic mean of precision and recall.

$$F1 - score = 2 * \frac{Recall * Precision}{Recall + Precision}$$

AUC

Area Under the Curve usually refers to the area under the precision-recall curve (Figure 2.3). A high AUC value implies a high-quality classification.

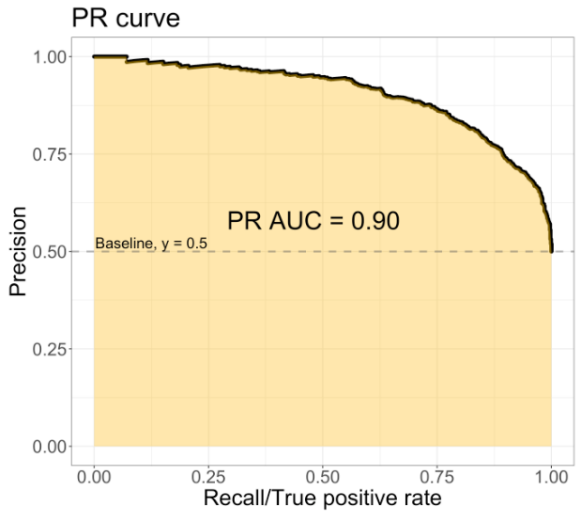


Figure 2.3 Precision-Recall Curve and AUC example

As explained above, the most important measure in our case is FN. Based on this, the order of priority of the metrics is decided to be accuracy, F1-score, sensitivity, and then the rest of them.

2.4 Monitoring Devices

In the series of experiments that are presented in Section 3, three monitoring devices were used in order to record the variance of the physiological data of the participants. These devices are BIOPAC MP150, Microsoft Band 2, and Moodmetric Smart Ring. BIOPAC was used in all four experiments, while the two other devices were used only in the fourth experiment. This section gives a brief explanation of each device and the signals they recorded.

2.4.1 BIOPAC MP150

BIOPAC MP150 [27] is a stationary device, which means it is installed in the Psychology lab and is primarily intended to operate there. It is a comprehensive tool that can record a wide range of signals, from 16 channels. It has a wide range of available sample rates, starting from 2 samples/hour and reaching 200 kHz [28]. BIOPAC was employed in the context of the research to record ECG, GSR, and fEMG (COR, ORB, and ZYG muscles) at a sampling rate of 1000Hz. It was used in combination with AcqKnowledge 3.9.0 [29] data acquisition software, which provides functionalities for quicker signal analysis.

2.4.2 Microsoft Band 2

Microsoft Band 2 [30] is a smart band that is worn on the wrist. It is a hands-free gadget that can be worn in everyday life. The sensors that were used in the fourth experiment are PPG and GSR. It can record PPG in 1Hz and GSR in 0.2/5Hz. As it can be observed, the available sampling rates are mush lower than the stationary unit, in order to save memory and power.

2.4.3 Moodmetric Smart Ring

Moodmetric Smart Ring [31] is a wearable device that can be worn on any finger of the hand. It a real-time device which can measure GSR in a sample rate of 3 Hz.

Chapter 3

Data Collection and Previous Work

3.1 Physiological Experiments 17

3.2 Related Work 19

The Department of Psychology of the University of Cyprus conducted a series of four experiments: Diagnosis of Experiential Avoidance in Smokers, Diagnosis of Eating Disorders, Diagnosis of Experimental Avoidance for Anxiety, and Functional versus Dysfunctional Coping with Acute Pain. The experiments took place in the psychology lab and the participants were volunteers. This section presents the procedure of each experiment, as well as the methodologies of works that are related to the topic of this thesis [1] [18] [32].

All experiments are related to Acceptance and Commitment Therapy, where people are split into two groups (acceptance or avoidance) based on their reactions. As explained in Section 1.2, acceptance-based strategies include individuals accepting their thoughts and sensations (functional). On the contrary, avoidance-based strategies are related to avoiding uncomfortable thoughts and sensations or attempting to control, alter or avoid them (dysfunctional) [33]. Also, one does not always fall into the same category; their classification changes depending on the environment and the circumstances. During the *data collection* of the first three experiments, it was assumed that each participant belonged to a single group throughout the whole procedure. In the case of the fourth group, this hypothesis did not apply.

3.1 Physiological Experiments

3.1.1 Diagnosis of Experiential Avoidance in Smokers

The aim of this experiment was to compare the ability of emotional regulation between smokers that belong to the category of acceptance and smokers that belong to the category of avoidance. The experiment was composed of five consecutive timeframes and each one had a duration of 8 minutes. The first timeframe was used as a baseline and to make sure that the participant was in a state of calm. In the two following timeframes, an emotionally neutral video was shown and during the two final timeframes the participant viewed a video that was supposed to evoke negative emotions. With the obtained data, an expert from the Department of Psychology classified the participant to one of the categories.

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

3.1.2 Diagnosis of Eating Disorders

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

With the obtained data, an expert from the Department of Psychology classified the participants to one of the categories. The signals that were recorded from the participants during the whole procedure were ECG, COR (using fEMG), and GSR with a sampling rate of 1000Hz. The participants were also asked to complete the *Body Image Acceptance and Action Questionnaire*

(BI-AAQ), which measures body image flexibility. Participants responded on a seven-point scale from *never true* to *always true*, where higher summed scores indicate greater body image flexibility [34].

3.1.3 Diagnosis of Experiential Avoidance for Anxiety

The aim of this experiment was to compare the ability of emotional regulation between people that belong to the category of acceptance and people that belong to the category of avoidance regarding anxiety. The experiment was composed of 72 consecutive timeframes and each one had a duration of around 1.8 minutes. In each timeframe the participant was showed a single image, that was supposed to cause a different reaction based on if they show signs of anxiety or not. With the obtained data, an expert from the Department of Psychology classified the participant to one of the categories.

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

3.1.4 Functional Versus Dysfunctional Coping with Acute Pain

In this study 80 people took part [35]. The aim was to compare acceptance and avoidance coping strategies in a pain-induction experiment. Participants were randomly split into 4 groups (conditions) and the participants of each group were given different instructions on how to deal with pain. The four conditions were: (a) Acceptance followed by avoidance; (b) Avoidance followed by acceptance; (c) No instructions given (control) followed by acceptance, and (d) No instructions given (control) followed by avoidance.

The experiment was composed of three timeframes. The first timeframe lasted 5 minutes, and was used as a baseline to make sure that the participant was in a state of calm. Afterwards, participants were instructed on how to deal with pain, in the following time frame, based on their condition. The second timeframe followed, in which participants were subjected to the Cold Pressor Task (CPT), which consists of each individual being asked to immerse their hand in a container filled with cold water for as long as they could. Following, participants were instructed on how to deal

with pain, in the last time frame, based on their condition. The third timeframe followed, in which participants were subjected to a second CPT. The maximum duration of the second and third timeframe was 3 minutes.

Multiple measures -- behavioural, psychophysiological and self-reported -- were recorded during the whole procedure. Behavioural measures are pain threshold and pain tolerance, which are the number of seconds that passed from immersion until the participant reported pain verbally and until the participant removed their hand from the container respectively. The psychophysiological signals collected using the stationary device are ECG, part of the EDA signal (SCL; explained in Section 4.2.2), and fEMG (COR and ZYG muscles), with sampling frequencies of 1kHz, 250Hz and 1kHz respectively. The measures collected using the band are PPG and EDA with sampling frequencies of 1Hz and 0.2Hz respectively. The only measure collected from the ring is EDA with a sampling frequency of 3Hz. Regarding self-reported data, participants completed some questionnaires that examined various aspects, including their psychological condition and their use of pain-coping strategies.

3.2 Previous Work

Three prior researches examined the data of the experiments regarding smoking, eating disorders, and anxiety and attempted to classify the participants using Machine Learning. The methodologies of the researches are presented in the current section.

This thesis focuses on the experiment that examines pain and emotions management and the work that has been done is presented in Chapters 4 to 6.

3.2.1 Diploma Project of Ch. Galazis in 2017

In the first analysis [1], in order to select the combination of features that yields the best results in the classification process for the experiments regarding smoking and eating disorders, knowledge from a previous research [2] was utilised. In regards to the experiment about anxiety additional work was done. All unique combinations of features were used to train and test the Random Forest classifier. The candidate features were the mean values of each recorded signal in each timeframe.

For each combination the accuracy of the resulted classifier was measured. The selected combination was the one with the highest accuracy and the smallest number of features.

The Machine Learning algorithms that were studied are: Logistic Regression, Naive Bayes, K-Nearest Neighbours, Classification Tree, Neural Network, SVM, Bagging (Decision Tree was used as the Base Learner), AdaBoosting (Decision Tree was used as the Base Learner), Gradient Tree Boosting, and Random Forest. The data were split into the training and test set, in ten different ways. Each algorithm was executed ten times, once for each split, and the results from the distribution that returned the best results were used in the comparison of the algorithms.

3.2.2 Master Thesis of A. Trigeorgi in 2018

In a more recent study [18], a different methodology was applied. In this study, more work was done in feature extraction, as time-domain features were extracted from ECG signal. Thus, the candidate features were not only the mean values of each signal, but also some time-domain features extracted from the ECG signal (explained in detail in Chapter 4.1.1). To select the best combination of features, Random Forest Classifier was used in conjunction with Stratified k-fold cross-validation. This means that the data were distributed in training and testing set in k different ways, for each combination of features Random Forest Classifier was executed k times and the average performance was measured.

The algorithms studied were the same as in the previous study. What was different in this study is the algorithms execution method, as in this study Stratified 5-fold cross-validation was used. The data were split into training and test set, in five different ways. Each algorithm was executed five times, once for each split, and the average performance of the five executions was measured.

3.2.3 Master Thesis of G. Demosthenous in 2019

This is the latest study [32] of the three studies examined. In this study, even more features were extracted from the ECG signal. To select the most effective combination of features, the method introduced by Breiman and Friedman was utilized [36] to calculate feature importance using

Gradient Boosting Decision Tree. This method ranks all the candidate features based on node impurity.

The algorithms studied were different from the two previous studies, as focus was given to algorithms that are based to trees. The five algorithms analysed were: Gradient Boosting Decision Tree (GBDT), Ada Boosting Decision Tree, Bagging Decision Tree (BDT), Random Forest (RF), and Extra Trees (ET). An extra step was done in this analysis, which was the training data multiplication, in order to increase the number of samples and negate the hypothesis that each participant belonged to the same group throughout the experiment. Two methodologies were used and their results were compared: Moving Window Methodology (MWM) and Rectangular Window Methodology (RWM). The algorithms execution method combined the methods used by the two previous studies. The data were split, to training and test set, in ten different ways using 10-fold cross-validation and each algorithm was executed 10 times for each split, leading to 100 executions per algorithm.

Chapter 4

Signal Analysis

4.1 Training Data Multiplication 22

4.2 Feature Extraction 24

4.3 Feature Selection 30

This chapter describes in detail the procedure used to handle the raw psychophysiological signals so that they can be used for algorithm training. As stated in Section 3.4.1, the signals recorded in the lab are ECG, part of EDA (SCL), and fEMG (from COR and ZYG muscles) using the stationary device. Regarding the wearable devices, using the band PPG and EDA were recorded, while using the ring only the EDA signal was recorded.

4.1 Training Data Multiplication

A major issue with the experiment that this thesis studies is the lack of training data. Data multiplication can be used to solve this problem. A related study [32] examined two data multiplication methodologies: *Moving Window Methodology* and *Rectangular Window Methodology*.

Moving Window Methodology

In Moving Window Methodology (MWM), a window is used which is moved along each signal recorded. As shown in Figure 4.1, starting from the beginning of each signal, the window moves

by 1 second and takes a chunk of data. Thus, a number of artificial samples are produced from each participant. These samples can then be handled as independent psychophysiological signals and labelled with the label of the signal from which they originate. As it can be observed, consecutive samples do not differ much, as they have a big part of data in common. As shown in previous research, this results in the overfitting of the model and the inability to generalise.

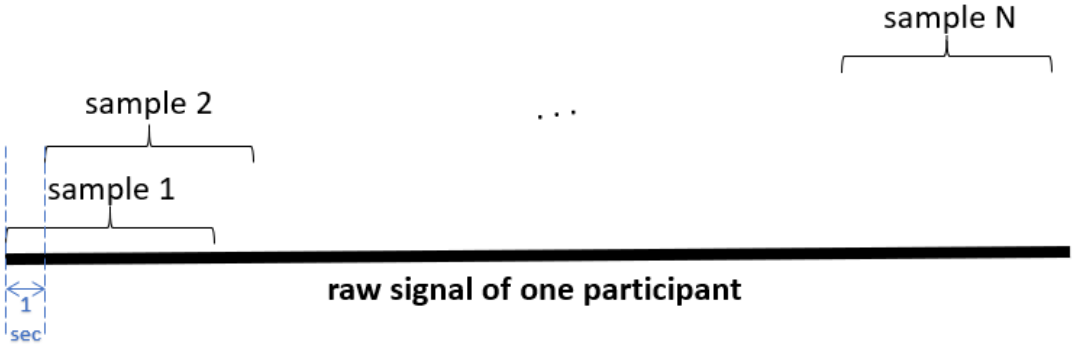


Figure 4.1 Moving Window Methodology [32]

Rectangular Window Methodology

Regarding Rectangular Window Methodology (RWM), the logic is very similar to the first with the difference that there is no overlapping between the windows (Figure 4.2Figure 4.2 Rectangular Window Methodology [13]). This is the methodology that has yielded the best results in previous research [32], and the one that will be used to multiply the data in the present work.

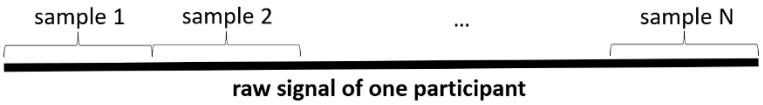


Figure 4.2 Rectangular Window Methodology [13]

Four different window sizes (10, 20, 30, and 40 seconds) were used to compare their efficiency in the classification phase. RWM was applied to all the recorded signals both from the wearables and the stationary devices. The total number of the generated samples is shown in Table 4.1. As it can be seen, in all window sizes the number of samples created is greater in the stationary devices than the wearables. A possible reason is the lack of synchronization of the moment when the two types of devices started recording. As mentioned in Section 3.4.1, 80 people took part in the experiment.

As it can be observed, the samples increased by 221-1413% in wearables devices and by 282-1688% in stationary devices.

	Total number of samples	
Window Size (sec)	Stationary Devices	Wearable Devices
10	1430	1210
20	693	590
30	451	386
40	306	257

Table 4.1 Number of artificial samples produced

4.2 Feature Extraction

The raw psychophysiological signals cannot be used to effectively train the algorithms. Thus, it is needed to extract important features from them. This section explains in detail the features extracted from each signal.

4.2.1 Features Extracted from ECG Signal

The ECG signal can be represented graphically and a basic pattern can be observed (Figure 4.3). In the order listed, the pattern is made up of three waves: P, QRS, and T. In essence, QRS is a wave complex, meaning that it consists of three waves: Q, R and S [7].

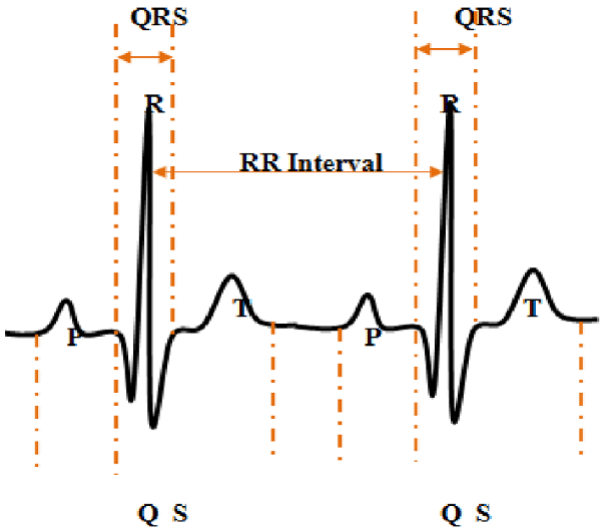


Figure 4.3 Visual representation of ECG signal [37]

There are three categories of features that can be extracted from the ECG signal: frequency-domain, spectral, and time-domain measures. Based on a previous study [18], the time-domain measures are the most sufficient in the topic of this thesis. Time-domain measures (also called HRV time-domain measures) are based on Heart Rate Variability, which is known as RR intervals. An RR interval is the time elapsed between two consecutive R peaks. RR intervals are also referred to as NN intervals. In fact, NN intervals are the time elapsed between two consecutive normal R peaks, which are the R peaks that do not include artifacts [38]. In practice, the two terms (RR intervals and NN intervals) are used interchangeably [39].

To extract the RR complex from the raw ECG signal, the BioSPPy Python toolbox was utilized. More specifically, using the method biosppy.signals.ecg.ecg the noise was removed and the indices of the R peaks were found [18]. Thereafter, RR intervals were computed, as shown in Table 4.2, as well as the differences (RRdiff) and squared differences (RRsqdiff) between consecutive RR intervals.

Abbreviation	Explanation	Formula
RR	As explained above, it is the interval of consecutive R waves in milliseconds.	$RR = \frac{diff(R\ peaks)}{sf} * 1000$, where sf is the sampling frequency
RRdiff	The absolute value of the differences between consecutive RR intervals	$RRdiff = \lvert diff(RR) \rvert$

Abbreviation	Explanation	Formula
RRsqdiff	The squared differences of consecutive RR intervals.	$RRsqdiff = RRdiff^2$

Table 4.2 Measures used to express time-domain features

Based on the above, the time-domain features of Table 4.3 can be derived [40] [41].

Feature Abbreviation	Feature Description	Formula	Unit
IBI	Inter-Beat Intervals. The average of RR intervals	$IBI = \overline{RR}$	ms
BPM	Beats Per Minute. The average number of heart beats per minute.	$BPM = \frac{60000}{\overline{RR}}$. As RR intervals are measured in milliseconds, to find bpm, it is needed to divide 60 000 ms (1 minute) with the mean of RR intervals.	bpm
SDNN	Standard Deviation of NN intervals	$SDNN = \sqrt{\frac{1}{N-1}\sum_{1=1}^{N}(RR_i - \overline{RR})^2}$	ms
SDSD	Standard Deviation of Successive Differences between consecutive RR intervals	$SDSD = \sqrt{\frac{1}{N-1}\sum_{1=1}^{N}(RRdiff_i - \overline{RRdiff})^2}$	ms
RMSSD	Root Mean Square of Successive RR interval Differences	$RMSSD = \sqrt{\frac{1}{N-1}\sum_{1=1}^{N}(RRdiff_i)^2}$	ms
pNN20	The ratio of differences of consecutive NN intervals that are greater than 20ms to	$pNN20 = \frac{count(diff(RR)>20\ ms)}{count(diff(RR))}$, where count(X) gives the number of elements in X	%

Feature Abbreviation	Feature Description	Formula	Unit
	all consecutive NN intervals		
pNN50	The ratio of differences of consecutive NN intervals that are greater than 50ms to all consecutive NN intervals	$pNN50 = \frac{count(diff(RR)>50\ ms)}{count(diff(RR))}$, where count(X) gives the number of elements in X	%
HRMAD	The Median Absolute Deviation of the Heart Rate	$HRmad = median(\lvert RR_i - \widetilde{RR} \rvert)$, where $\widetilde{RR} = median\ (RR)$	bpm

Table 4.3 Time-domain features extracted from ECG

4.2.2 Features Extracted from EDA Signal

There are four types of features that can be extracted from EDA. These are time domain features, frequency domain features, time-frequency domain features, and Mel-frequency cepstrum features [42]. This thesis focuses on time domain features and more specifically in event-related features, because features of this type had efficient results in a previous study [42], and statistical features, as they are widely used in the literature. To extract the features explained below from the raw EDA signal, the NeuroKit2 Python toolbox [43] was used. More specifically, using the method neurokit2.eda_process the signal was cleaned from noise and the needed measurements for computing the features were found.

Event-related Features

The EDA complex comprises two components: Skin Conductance Level (SCL) and Skin Conductance Response (SCR). SCL refers to tonic changes, which are changes in the overall baseline of EDA activity, meaning that it fluctuates slightly on a time scale ranging from tens of

seconds to minutes. SCR, on the other hand, corresponds to phasic changes, which are rapid fluctuations over time that usually occur after the beginning of a stimulus [5] [6]. While the tonic level is insufficiently informative, SCR is riding on top of it (Figure 4.4) and has rapid changes and easy-to-spot peaks that can be helpful in recognising emotional stimulus events [44]. Thus, many SCR-related features, also called event-related features, can be extracted from the EDA complex.

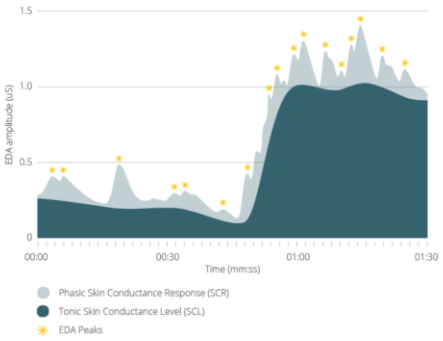


Figure 4.4 Visual representation of EDA signal and its components [45]

Figure 4.5 illustrates some of the quantitative measurements that will be used to extract features from the SCR signal. When a person is exposed to a certain stimulus, their initial level of SCR signal rises rapidly, reaches a peak and then drops slowly. Latency corresponds to the time elapsed from stimulus exposure to the onset of the SCR burst, and ranges between 1-5 seconds. Peak amplitude is the difference in amplitude between the onset and the peak [44]. To reject minor changes in the signal that cannot be classified as SCR peaks, a threshold of 0.05 μS is usually used [46]. Rise time is the duration from the onset to the peak. [44].

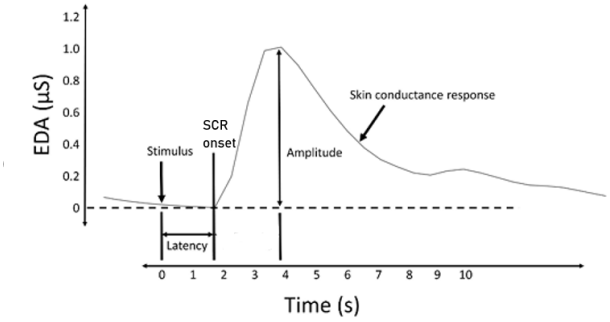


Figure 4.5 Graphical representation of the main SCR components [47]

The features of Table 4.4 were extracted from the SCR signal [42] within a window of 3 minutes, which corresponds to the maximum duration the participants were allowed to have their hand in the water.

Feature Name	Feature Description
Peak Count	The number of peaks in the window
Mean Peak Amplitude	The average value of amplitudes in the window
Sum Peak Amplitude	The summation of amplitudes in the window
AUC	Area under the curve of the phasic component

Table 4.4 Event-related features extracted from EDA

Statistical Features

Statistical features include mean value, standard deviation, kurtosis and skewness of the signal [42]. The statistical features extracted can be seen in Table 4.5.

Feature Name	Feature Description
Mean EDA	Average value of EDA
Standard Deviation EDA	Standard Deviation of EDA
Kurtosis EDA	Kurtosis of EDA
Skewness EDA	Skewness of EDA
Mean SCL	Average value of SCL
Standard Deviation SCL	Standard Deviation of SCL
Mean SCR	Average value of SCR
Standard Deviation SCR	Standard Deviation of SCR

Table 4.5 Statistical features extracted from EDA

It is worth mentioning that in some cases, there were no healthy SCR peaks detected in the EDA signal that was recorded by the ring. Therefore, it was not possible to extract features from that signal.

4.2.3 Features Extracted from fEMG Signal

Firstly, both COR and ZYG signal, were processed to remove noise and artefacts. This was done using the Python toolbox NeuroKit2 [43]. Afterwards, mean value, root mean square value, median value and mean absolute value of each signal were extracted, as they brought good results in a related work [48].

4.3 Feature Selection

If we use all the extracted features to train the classifiers, this will greatly increase the time complexity of the algorithms and require a lot of memory resources. Also, the performance of the algorithms may be reduced, due to overfitting or in case there are features that are not related to the target. For this reason, we perform feature selection so that we can select the features that appear to have better performance when used in the classification.

When it comes to supervised data, there are numerous feature selection approaches to choose from. More specifically, there are three major categories: Wrapper, Filter, and Embedded Methods. Hybrid Methods, which are a mix of filter and wrapper methods are also used sometimes [49]. The three main categories of methods mentioned will be examined, in order to compare their results. Figure 4.6 shows the main concept behind the three main categories, which are explained in detail in this chapter.

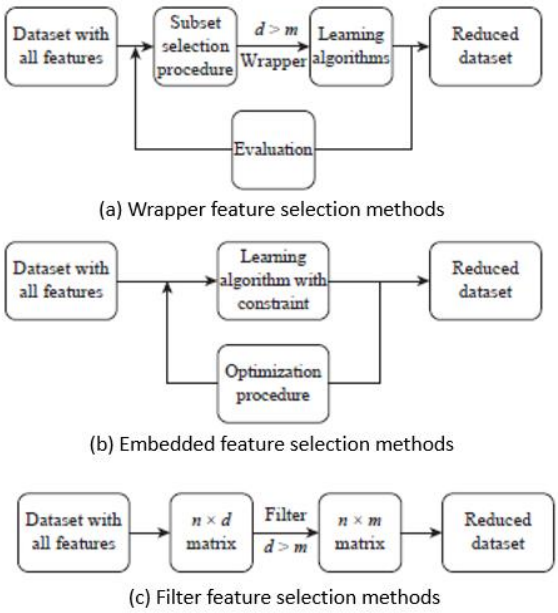


Figure 4.6 Flowchart of the three main feature selection methods [50]

In this subsection, the three main categories of feature selection methods are explained, and one technique from each category is used to perform feature selection on the available data.

4.3.1 Wrapper Methods

4.3.1.1 Technique

In wrapper methods, a certain machine learning algorithm is used, which we attempt to fit on the data. By searching the whole space of possible features, these methods try to find the feature combination that yields the best performance. Several performance measures can be used. In the case of classification, metrics such as accuracy, precision, recall, and f1-score are the most common. Wrapper Methods have high computation time, especially when there are many possible features in the dataset [51].

The main feature selection techniques that are under wrapper methods, are Forward Selection, Backward Elimination and Exhaustive Feature Selection. In Forward Selection starting with the best performing feature, in each iteration another variable is chosen that gives the best performance when combined with the already selected features. This process continues until the predetermined performance requirement is met. Backward Elimination is the reverse of Forward Selection, as first we fit the model using all the possible features, and in each iteration a feature is removed,

based on p-value. A p-value is a statistical measure that is used to validate a null hypothesis against observed data. The null hypothesis is the hypothesis that the observed difference between specified populations is caused by to errors and it is not statistically significant. Exhaustive Feature Selection performs a brute-force search, meaning that every potential subset of features is examined and the one with the highest performance is selected [49]. This thesis examines the Exhaustive Feature Selection technique, in order to compare its results with a related work [2].

4.3.1.2 Results

The Machine Learning algorithm used is the Random Forest Classifier with 1000 estimators, to be consistent with previous work. The evaluation criterion was the average accuracy of a feature combination between ten executions. In each execution the data were split into different class-balanced training and test sets, meaning that 2/3 of the samples that belong to the acceptance category and 2/3 of the samples that belongs to the avoidance category were used for training and the rest samples were used for testing.

Since there is a large number of possible features (18 in stationary devices, and 33 in wearable devices), it was decided to examine the performance of all the feature combinations of size smaller than four. Anyhow, a feature combination with more than three features would cause time, memory, and performance issues, as previously explained. The total number of feature combinations that were examined are 987 and 6017 for the stationary and wearable devices' data respectively.
The execution time of feature selection on stationary devices was three hours for all the four window sizes, whereas it took around four hours for each window size on wearable devices, which means about twelve hours in total. This demonstrates that the execution time of wrapper methods is notably long, especially when dealing with high dimensional data.

Table 4.6 shows, in descending order, the three best-performing combinations of features for the three window sizes and the two types of devices. As we can see, relatively good accuracy has been achieved in all cases. Also, both in stationary and in wearable devices, the results are quite similar between the three different window sizes, although the accuracy drops as the window size increases. This was expected, because as the window size increases the number of generated samples decreases. Thus, the purview of the algorithm is limited and the accuracy drops. Moreover,

in most of the cases, the feature combinations that brought the best results were of size three, while in a few cases two-feature combinations also brought good results.

With regard to stationary devices, the combinations that brought the best results include features mainly from the EDA signal (i.e., sweat) and the COR and ZYG signals (i.e., face muscles). Specifically, from the EDA signal, the average of the tonic level (SCL) brought the best results and from the other two signals the Mean Absolute Deviation and the Root Mean Square.
Regarding wearable devices, there are various combinations of features that brought good results. Most of these include the average EDA signal recorded from the ring, the average amplitude of the SCR signal recorded from the band, and the average heart rate.

It is noteworthy that in both types of devices, the best combinations include features extracted from the EDA signal. Nevertheless, on stationary devices this is the SCL while on wearables, which had a wider range of features extracted from EDA, other features seemed more popular in the combinations.

<table>
<tr><th>Window Size</th><th>Data from Stationary Device (three best feature combinations)</th><th>Data from Wearables Devices (three best feature combinations)</th></tr>
<tr><td>10</td><td>
<table>
<tr><th>Combination</th><th>Accuracy</th></tr>
<tr><td>EDA_meanSCL, COR_mad, ZYG_rms</td><td>90.85%</td></tr>
<tr><td>EDA_meanSCL, COR_rms, ZYG_rms</td><td>90.42%</td></tr>
<tr><td>EDA_meanSCL, COR_mad, ZYG_mad</td><td>90.21%</td></tr>
</table>
</td><td>
<table>
<tr><th>Combination</th><th>Accuracy</th></tr>
<tr><td>ECG_hrmean, EDAwatch_meanSCL, EDAring_mean</td><td>89.77%</td></tr>
<tr><td>ECG_hrmean, EDAwatch_mean, EDAring_mean</td><td>88.52%</td></tr>
<tr><td>ECG_ibi, EDAwatch_mean, EDAring_mean</td><td>88.27%</td></tr>
</table>
</td></tr>
<tr><td>20</td><td>
<table>
<tr><th>Combination</th><th>Accuracy</th></tr>
<tr><td>EDA_meanSCL, COR_mad, ZYG_mad</td><td>90.74%</td></tr>
</table>
</td><td>
<table>
<tr><th>Combination</th><th>Accuracy</th></tr>
<tr><td>ECG_hrmean, EDAwatch_meanPeakAmp, EDAwatch_std</td><td>86.67%</td></tr>
</table>
</td></tr>
</table>

Window Size	Data from Stationary Device (three best feature combinations)		Data from Wearables Devices (three best feature combinations)	
	EDA_meanSCL, COR_mad, ZYG_rms	89.42%	EDAwatch_meanPeakAmp, EDAwatch_stdSCR, EDAring_mean	86.67%
	EDA_meanSCL, COR_rms, ZYG_mad	88.54%	ECG_ibi, EDAwatch_meanPeakAmp, EDAring_mean	86.15%
30	Combination	Accuracy	Combination	Accuracy
	EDA_meanSCL, COR_mad, ZYG_mad	87.16%	ECG_hrmean, EDAwatch_meanSCL, EDAring_mean	84.37%
	EDA_meanSCL, COR_rms	86.48%	EDAwatch_meanPeakAmp, EDAwatch_meanSCL, EDAring_mean	83.59%
	COR_mad, ZYG_mad	86.48%	ECG_hrmean, EDAwatch_meanPeakAmp, EDAwatch_mean	82.81%
40	Combination	Accuracy	Combination	Accuracy
	COR_mad, ZYG_rms	86.00%	ECG_sdsd, ECG_hr_mean, EDAring_stdSCR	75.29%
	COR_mean, COR_mad, ZYG_rms	86.00%	ECG_bpm, ECG_hr_mad, EDAwatch_mean	74.11%
	COR_mad, ZYG_mean, ZYG_rms	86.00%	ECG_sdnn, ECG_hr_mean, EDAring_skew	74.11%

Table 4.6 Results of Feature Selection using Wrapper Method

4.3.2 Embedded Methods

4.3.2.1 Technique

Embedded methods use classification algorithms that have built-in feature selection functionality [50]. There are two types of embedded methods that are widely used. Techniques that belong to the first type utilize inherent characteristics of decision tree algorithms, such as Random Forest and Classification and Regression Tree (CART). The second type involves regression l1 regularization using least absolute shrinkage and selection operator (LASSO) [50].

This thesis makes use of the first type, to be consistent with previous work where the results were excellent. The methodology used is the one that was proposed by Breiman and Friedman [36]. Their methodology uses CART and computes feature importance based on Gini impurity [52]. Gini impurity, also known as Mean Decrease Impurity, is defined as the overall decrease in node impurity (a measure of the homogeneity of the labels at the node) weighted by the probability of reaching that node [53]. The more a feature reduces impurity, the more significant it is [54].

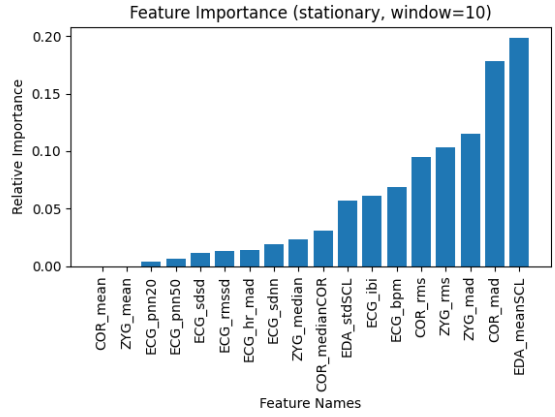
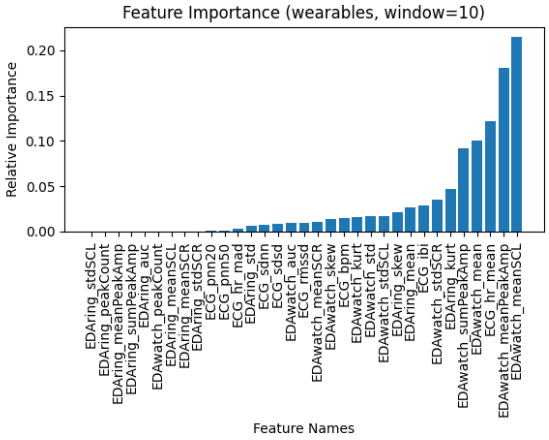
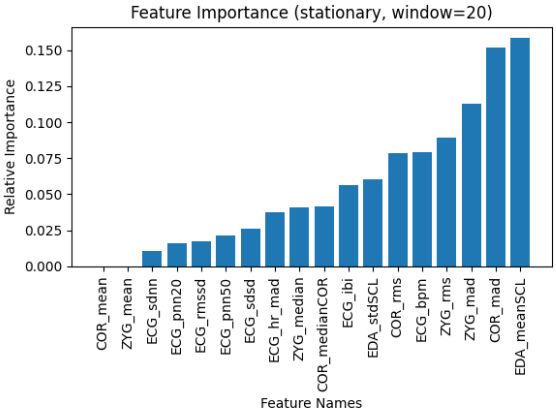
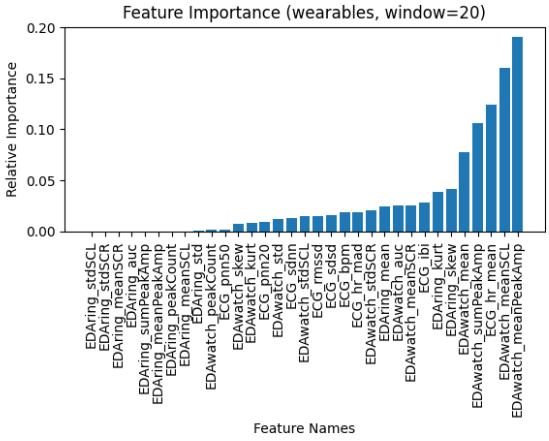
4.3.2.2 Results

In each of the eight cases, the importance of each feature has been normalized so that the sum of all the weights sums up to one. Table 4.7 shows the importance of the features in all cases.
As we can see, the average tonic level is the most important in stationary devices across the three smaller window sizes (SCL). The features that follow are the statistical properties of the COR and ZYG signals, and more specifically MAD. Moreover, in the 40-second window, inter-beat intervals (ECG_ibi) appear to be the most important feature.

Regarding wearable devices, the five most important features in all three window sizes are statistical features from the EDA signal recorded by the band (namely the average of the entire signal and its tonic level), time domain features from the EDA signal recorded by the band (sum and average of the amplitude of the phasic signal) and the average of heart rate. Of these, the most important appear to be the average amplitude and the average tonic level. Finally, while the same

signal (EDA) was recorded on both the band and the ring, only the features extracted from the band had good results. More about the comparison of the devices can be found in Section 6.

As we can observe, the average tonic level brought good results in both stationary and wearable devices. The other features that had good performance are not the same, and cannot be compared, as the signals they were extracted from were not recorded in both device types.

	Data from Stationary Device	Data from Wearable Devices
10 seconds window		
20 seconds window		

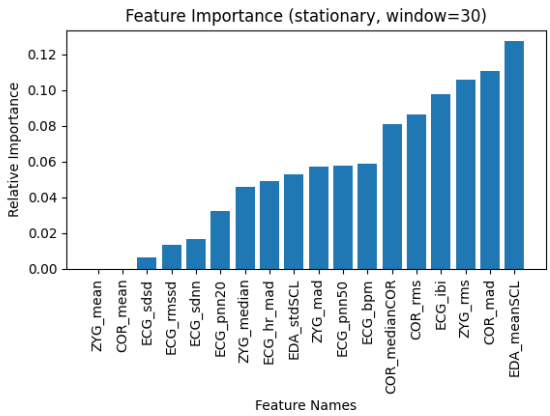
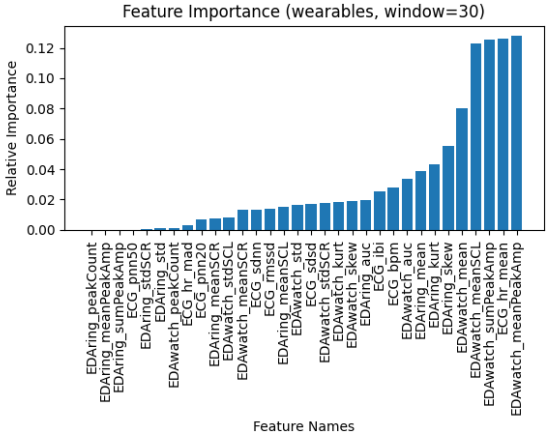
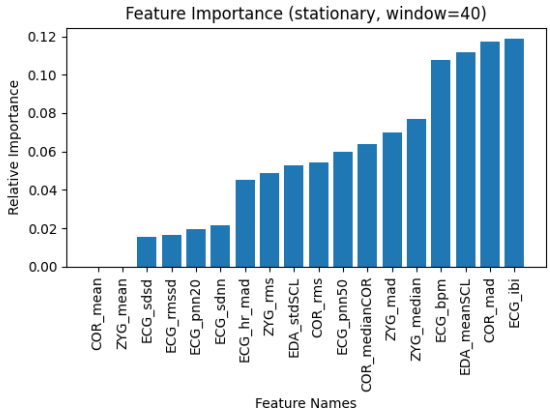
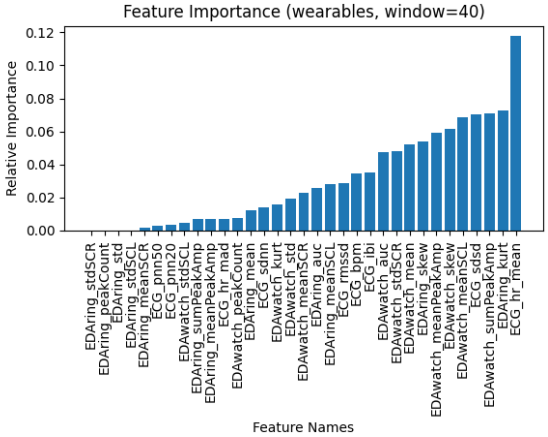
	Data from Stationary Device	Data from Wearable Devices
30 seconds window		
40 seconds window		

Table 4.7 Results of Feature Selection using Embedded Method

4.3.3 Filter Methods

4.3.3.1 Technique

Unlike the previous two methods, filter methods are not related to any Machine Learning algorithm. On the contrary, they get as an input the candidate features, they filter them and return those that are more related to the target variable. I believe this is the first time this technique is used in this line of work. There are various techniques that belong to filter methods and the most popular ones are Pearson Correlation, Chi-square test, and Information Gain. The choice of method to be used depends on various factors such as whether the features and the target variable are numerical or

categorical values [55] [56]. The main advantages of filtering methods are that due to the fact that they do not rely on a specific algorithm, the selected features are "generic, having incorporated few assumptions" [57]. Additionally, they usually need less computational resources and have a fast execution time.

The technique that this thesis utilizes is Pearson correlation coefficient, as it gives information about both the magnitude (i.e., strength) of the correlation and the direction of the relationship. Pearson correlation ranges between -1 and +1, with -1 indicating total negative linear correlation, 0 indicating no connection, and +1 indicating total positive correlation. The correlation between two variables, x and y, is commonly represented as $\boldsymbol{r_{xy}}$ and is defined as [58]:

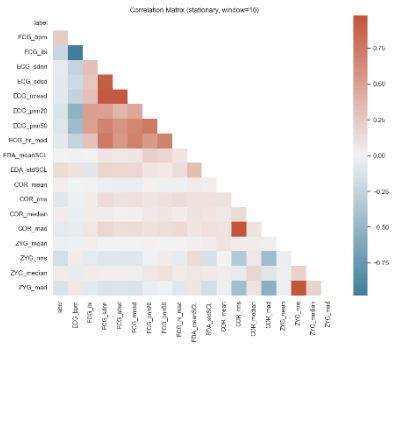
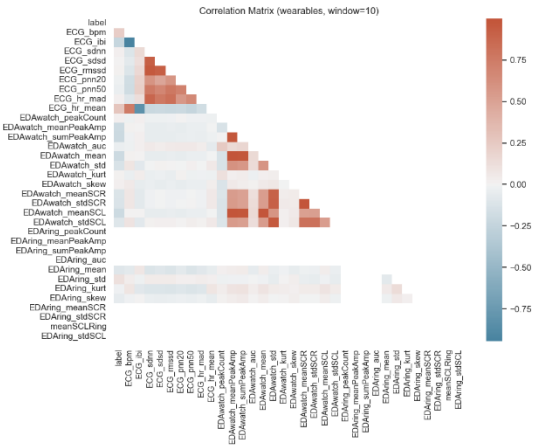
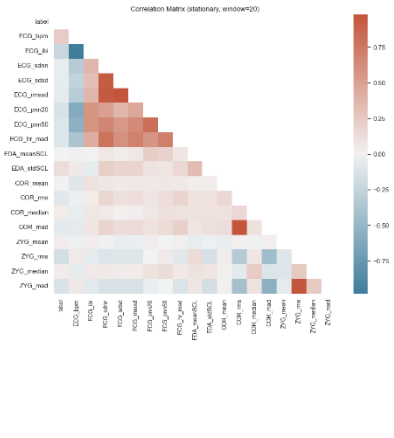
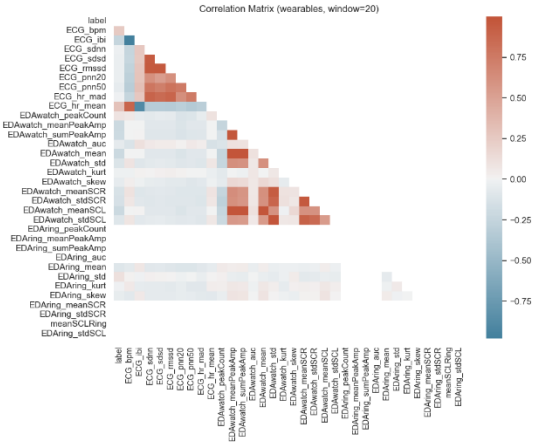
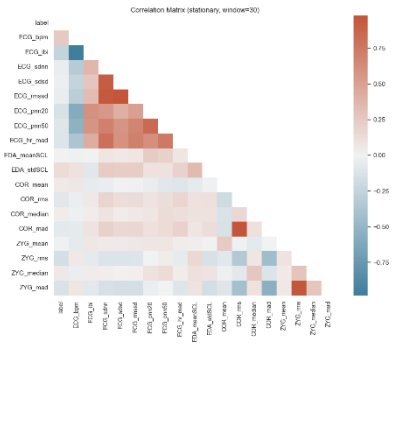
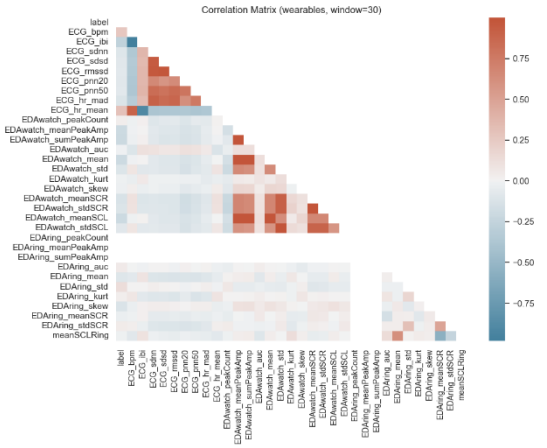
$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where

- n is the sample size of the dataset
- x_i, is the value of feature *x* in the *i*-th sample; and analogously for y_i
- $\bar{x}$ is the mean of the feature *x* in all data samples; and analogously for $\bar{y}$

4.3.3.2 Results

We can compute Pearson correlation between all features and the target variable, as well as between all possible pairs of features. These correlations can be represented in a matrix, known as correlation matrix. To do this, Seaborn Python library [59] was used. The results are shown in Table 4.8. Only the lower triangle of the matrix is shown as it is symmetrical. As shown in the legend, red color corresponds to positive correlation, blue color to negative correlation, while dark tones represent high correlations and lighter tones lower correlations.

	Data from Stationary Device	Data from Wearable Devices
10 seconds window		
20 seconds window		
30 seconds window		

40 seconds window

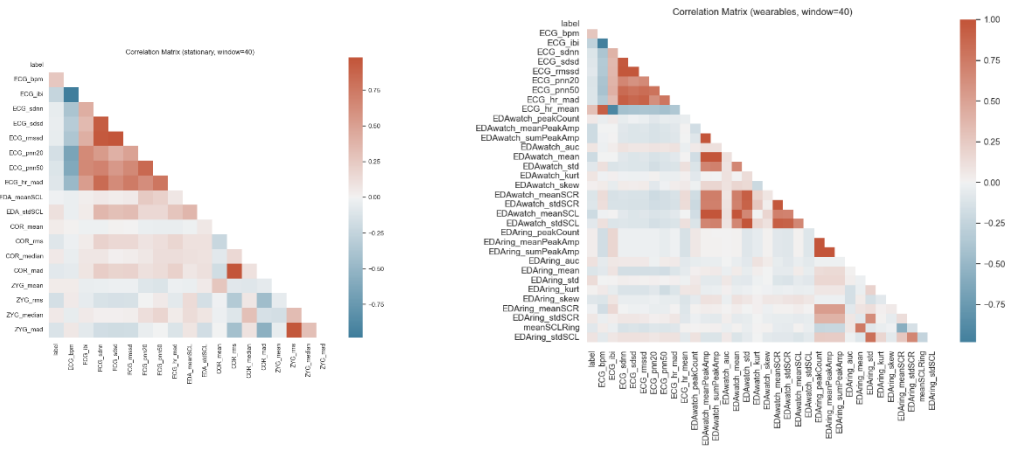


Table 4.8 Results of Feature Selection using Filter Method

It is clear that the correlations between the different window sizes in the two types of devices are nearly identical. The gaps shown in the three smaller window sizes in wearable devices were caused because there were no healthy SCR peaks detected in the EDA signal of the ring in the time period corresponding to the window size, because instead the instant EDA signal was recorded instead of the raw EDA (this was fixed in the next experimental phase). For this reason, the case of the 40-second window was also analyzed. What is important is that none of the features seems to be highly correlated with the target variable ('label' column), with the pulse value having the highest correlation with the target variable and reaching 0.3. We can also notice that in both devices' types there is a negative correlation between the pulse value (ECG_bpm, ECG_hr_mean) and the other features derived from the ECG signal, while the other features that are extracted from the ECG signal appear to have a fairly high negative correlation.

Regarding stationary devices, there appears to be a high negative correlation between RMS and MAD features in both muscles (COR and ZYG). In wearable devices we observe that the features derived from the band are positively correlated with each other, as well as the statistical characteristics of the SCR, SCL and EDA. There is also a high positive correlation between the static characteristics of STD, SCL, and EDA with the features that are related to peak amplitude. The other features have a very low correlation with each other.

4.3.4 Conclusion

Comparison of the three feature selection methods

Wrapper and embedded methods appear to have quite similar results in both types of devices. Based on the results of these two methods, the most effective features of the stationary devices are the average value of SCL and statistical properties of COR and ZYG signals. Regarding wearable devices, the average value of EDA signal, SCR amplitude, and heart rate have the best performance. By comparing the results of the wrapper methods' technique and the filter methods' technique, we can observe that features that belong in combinations that brought good results in the wrapper method have a very high correlation in the filter method. A possible reason is that all the features that belong in a certain combination have great results, and adding them all together makes the combination a little more efficient and therefore brings it higher in the ranking. However, if we use all those features, the training phase will take more time. Thus, the filter method is important to detect redundant features.

Regarding the comparison of embedded and filter methods, among all the features that seemed to have high feature importance in the embedded method, only the average value of the heart rate is related to the target variable, based on the correlation matrix.

Feature selection based on the results of the three methods

As we want to investigate the classification accuracy of wearable devices, we will avoid using features extracted from stationary devices' signals. One of the features that yielded good results in most of cases (both in stationary and wearable devices) is the average value of the tonic level of the EDA signal (SCL). However, based on the selection methods that were performed in data from the wearable devices, mean peak amplitude (which was not available in stationary devices) had a very similar performance to the average value of SCL. Also, mean peak amplitude is considered more reliable in the bibliography, because SCL varies at different levels depending on the person. Thus, the first feature we choose is ***mean peak amplitude***.

In addition, we could use the average value of the EDA signal as well as the average value of SCR amplitude. However, when looking at the correlation matrix, we can notice that the average value

of SCL has a strong correlation with these two features, so they would be redundant. The average value of heart rate had really good performance in both wrapper and embedded methods in stationary devices. Despite the fact it did not appear in any combination in the wrapper method in stationary devices analysis, in embedded methods it had a relatively high feature importance and it was not low in the ranking. Therefore, the ***average value of heart rate*** is the second feature we choose.

From the correlation matrix, it can be observed that any other feature has a strong correlation either with mean peak amplitude or heart rate. Hence, we will keep only these two features, since as explained at the beginning of the current subsection, smaller feature sets are preferable.

As an extra step, a representation of the data distribution can be performed. In Figure 4.7 a 2-D distribution of data with regard to mean peak amplitude and heart rate is shown. Each axis corresponds to a feature, and the colour of dots distinguishes the two groups. It is obvious that there is a big difference between the number of samples of the two groups, as most samples belong to the functional group. Also, the two groups are not separated in two distinct areas in space, as one would expect. Most of the data are between the values 0 – 2 μS of mean peak amplitude, and possibly the few that are left out correspond to noise that the tool did not manage to remove.

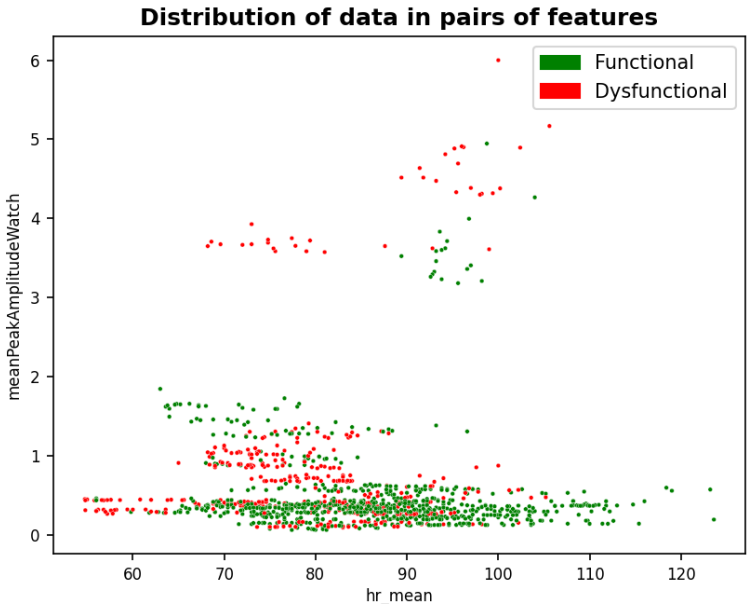


Figure 4.7 Distribution of data based on the selected features

A closer view of the dense area is shown in Figure 4.8. Although both groups have most of their values in the range 0-0.75μS of mean amplitude, and 55-100 bpm of heart rate, there is a differentiation in the rest values. Most samples of the dysfunctional group have amplitude values between 0.75-1.25 μS, while only the functional group's samples have amplitude higher than 1.45 μS and heart rate higher than 95 bpm.

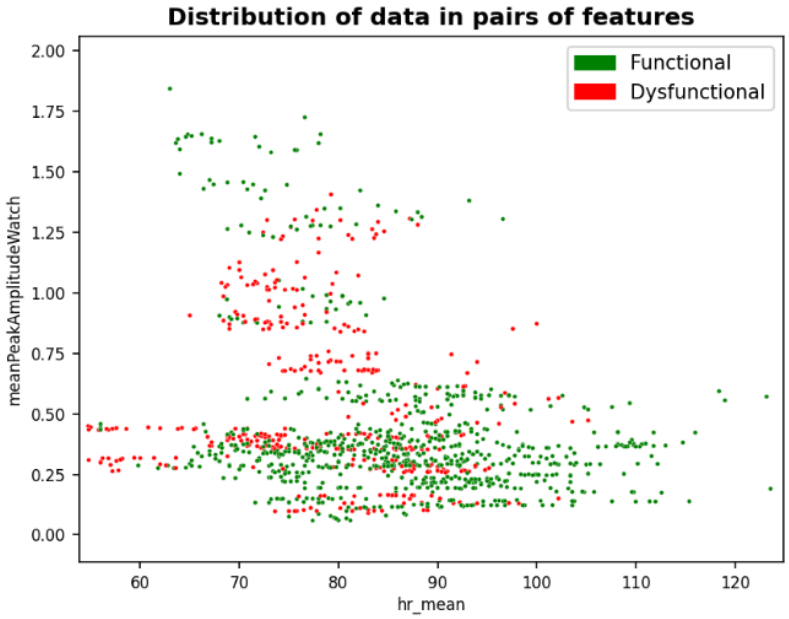


Figure 4.8 Distribution of data based on the selected features (zoomed in the dense area)

To confirm that more features would be redundant for the classification, a 3-D visualization of the data can be performed. The third feature used is the average value of EDA signal, as the combination of these three features has a good performance based on the wrapper method. Data distribution from two different viewpoints can be seen in Figure 4.9 and Figure 4.10. From the first viewpoint, it is noticeable that samples are still concentrated in a specific area. The second viewpoint verifies the linear correlation between mean peak amplitude and mean of EDA signal, that was indicated in the correlation matrices.

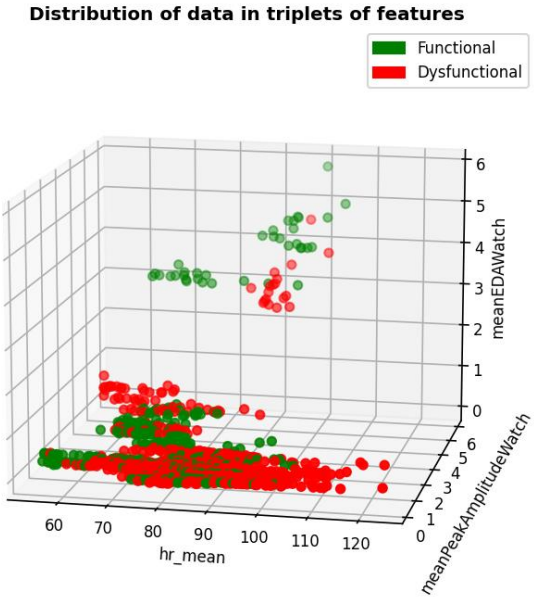


Figure 4.9 3-D Distribution of data from viewpoint 1

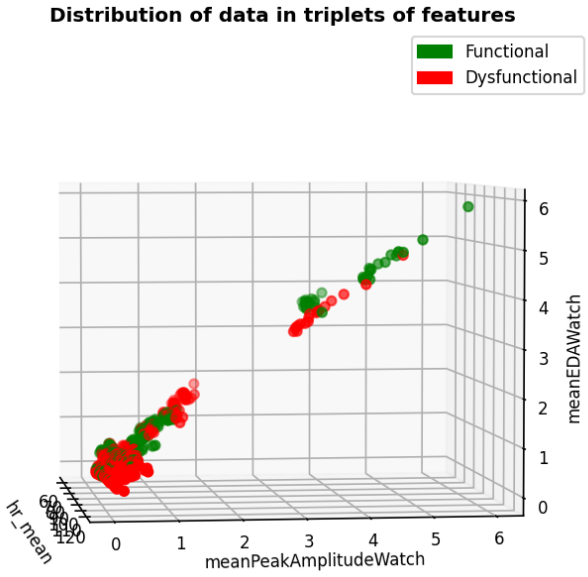


Figure 4.10 3-D Distribution of data from viewpoint 2

Chapter 5

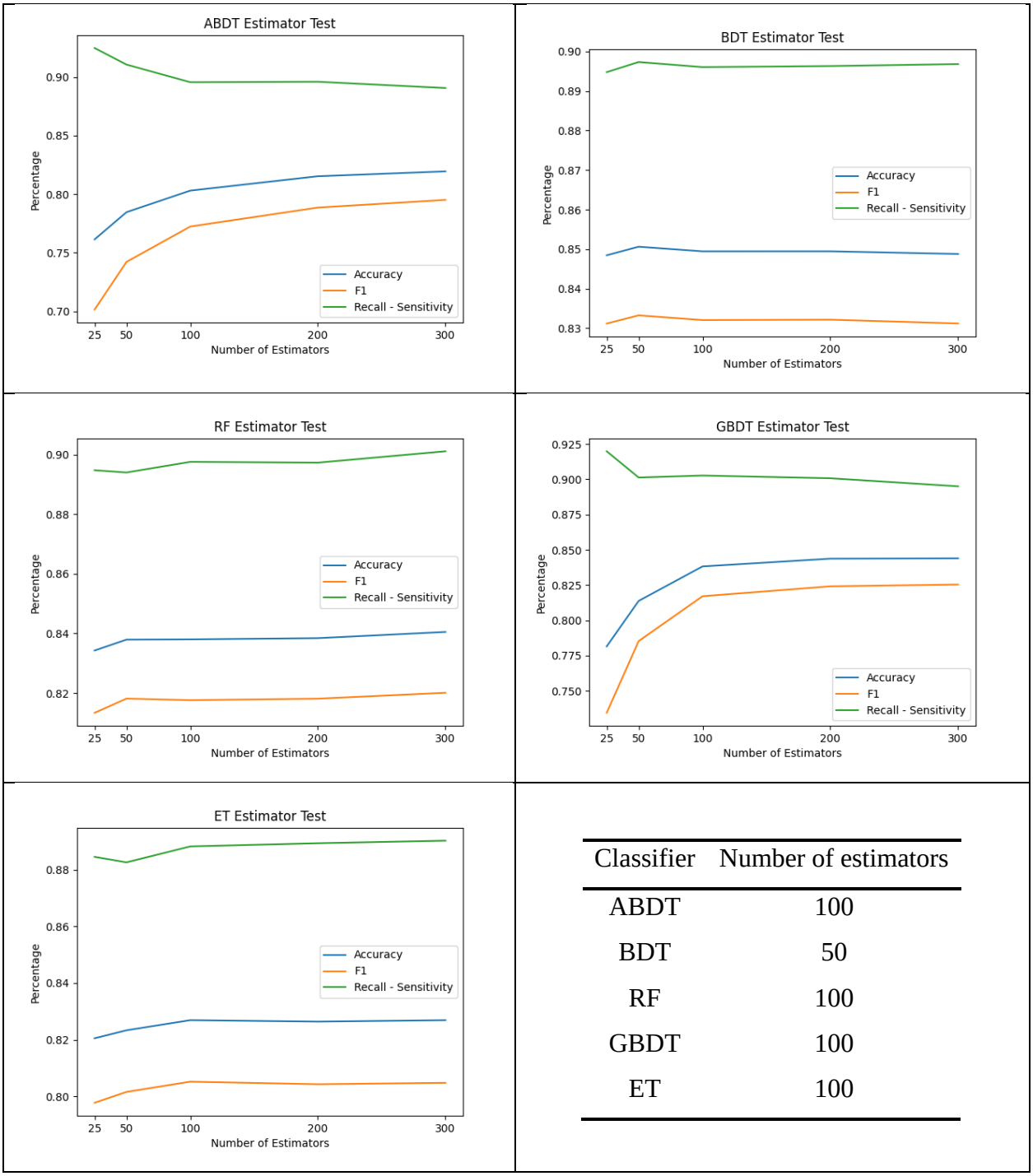
Classification

5.1 Classifier Fine-Tuning 45

5.2 Classifier Selection 47

5.1 Classifier Fine-Tuning

An important step before the training of the classifiers is to tune their parameters, in order to perform in the best way possible on the available data. The performance of the five classifiers upon several numbers of decision trees (known as estimators) will be examined. The three performance metrics are accuracy, F1- score and recall. The window size for all the experiments will remain constant at 20 seconds, and five numbers of estimators will be tested: 25, 50, 100, 200 and 300. The window size was chosen to be 20 seconds, to keep a balance between the number of data and the processing time. All classifiers will be trained using only the two features selected in the previous section (average value of SCR peak amplitude and average value of heart rate). Although more estimators usually result in better classification performance, they also need more processing time. Thus, this a serious trade-off that needs to be examined is the increase in performance and in processing time. The results are shown in Figure 5.1.



Classifier	Number of estimators
ABDT	100
BDT	50
RF	100
GBDT	100
ET	100

Figure 5.1 Classifier fine-tuning results

BDT, RF, and ET have an increase in their performance, as the number of estimators is rising, end after a certain value the performance changes negligibly. Hence, 50, 100, and 100 number of estimators were chosen, respectively. In ABDT and GBDT classifiers' accuracy and F1-score rise, while recall drops. The number of estimators chosen for both of them is 100, in order to achieve a balance between the values of the three metrics.

5.2 Classifier Selection

The next, and final, step is to compare the five classifiers. The performance metrics used are those explained in Section 2.3.2. Due to the fact that the experiment analysed is related to health, special importance was given, in order of priority, to accuracy, F1-score, and sensitivity. The evaluation methodology used is Stratified 10-fold cross-validation. This means that data were split in 10 different ways in training and test sets and afterwards each algorithm was executed 10 times for each split and its average performance on the seven metrics was calculated. Thus, each classifier was executed 100 times. This was done in all four window sizes. The value for each metric along with the corresponding standard deviations are shown in Table 5.1.

10 seconds window										
	GBDT		ABDT		BDT		RF		ET	
	Value	SD	Value	SD	Value	SD	Value	SD	Value	SD
Accuracy	0.84	0.03	0.8	0.03	0.85	0.03	0.84	0.03	0.83	0.03
F1-score	0.82	0.04	0.77	0.04	0.83	0.03	0.82	0.03	0.8	0.04
Sensitivity	0.9	0.04	0.9	0.04	0.89	0.03	0.9	0.03	0.89	0.03
Precision (PPV)	0.86	0.03	0.82	0.03	0.88	0.03	0.86	0.03	0.85	0.03
AUC	0.9	0.03	0.87	0.03	0.91	0.02	0.91	0.03	0.89	0.03
Specificity	0.72	0.07	0.63	0.07	0.77	0.06	0.73	0.06	0.71	0.07
NPV	0.8	0.07	0.77	0.07	0.8	0.05	0.79	0.05	0.78	0.06
20 seconds window										
	GBDT		ABDT		BDT		RF		ET	
	Value	SD	Value	SD	Value	SD	Value	SD	Value	SD
Accuracy	0.82	0.05	0.79	0.05	0.82	0.05	0.82	0.05	0.81	0.05
F1-score	0.79	0.06	0.76	0.05	0.8	0.06	0.8	0.05	0.79	0.05
Sensitivity	0.89	0.05	0.88	0.06	0.88	0.06	0.88	0.05	0.88	0.05
Precision (PPV)	0.84	0.04	0.82	0.04	0.86	0.04	0.85	0.04	0.84	0.04
AUC	0.88	0.05	0.85	0.05	0.88	0.05	0.87	0.05	0.87	0.05
Specificity	0.68	0.1	0.63	0.11	0.72	0.1	0.71	0.09	0.69	0.09

NPV	0.78	0.09	0.75	0.09	0.77	0.09	0.77	0.08	0.76	0.08
30 seconds window										
	GBDT		ABDT		BDT		RF		ET	
	Value	SD	Value	SD	Value	SD	Value	SD	Value	SD
Accuracy	0.79	0.05	0.79	0.06	0.81	0.05	0.81	0.06	0.78	0.03
F1-score	0.76	0.07	0.76	0.07	0.79	0.06	0.78	0.06	0.75	0.04
Sensitivity	0.87	0.06	0.86	0.07	0.85	0.07	0.86	0.07	0.86	0.04
Precision (PPV)	0.82	0.05	0.82	0.05	0.86	0.05	0.85	0.05	0.81	0.03
AUC	0.85	0.06	0.83	0.07	0.87	0.06	0.88	0.06	0.83	0.03
Specificity	0.63	0.14	0.65	0.12	0.73	0.12	0.71	0.12	0.62	0.06
NPV	0.74	0.1	0.73	0.11	0.74	0.09	0.74	0.1	0.71	0.06
40 seconds window										
	GBDT		ABDT		BDT		RF		ET	
	Value	SD	Value	SD	Value	SD	Value	SD	Value	SD
Accuracy	0.75	0.07	0.76	0.07	0.74	0.08	0.74	0.07	0.73	0.08
F1-score	0.71	0.08	0.73	0.08	0.7	0.09	0.7	0.08	0.69	0.1
Sensitivity	0.84	0.09	0.83	0.09	0.83	0.1	0.84	0.09	0.83	0.1
Precision (PPV)	0.79	0.06	0.81	0.06	0.79	0.06	0.78	0.06	0.78	0.07
AUC	0.8	0.08	0.8	0.09	0.8	0.08	0.8	0.07	0.8	0.08
Specificity	0.57	0.15	0.63	0.15	0.58	0.15	0.56	0.16	0.54	0.17
NPV	0.68	0.14	0.68	0.13	0.67	0.15	0.67	0.13	0.66	0.15

Table 5.1 Results of all classifiers in all window sizes

The overall results are very good. The most effective algorithm in each window size is coloured differently. It can be observed that as the window size increases, the effective of all the algorithms drops. Hence, the best results are yielded in the data produced using the 10-second window. The performances of the five classifiers within a single window size are very close to each other. The best algorithm is ***Bagging Decision Tree***, with an excellent average accuracy of 85%. It is ideal that accuracy is in the range 80-90%, as this suggests that the classification is of high quality and that the model's chances of overfitting are low. Moreover, the standard deviation of all metrics is low, indicating that there is low variance in the efficiency of the model.

Chapter 6

Comparisons

6.1 Comparison of Psychophysiological Features Between Devices 49

6.2 Comparison of Results with Previous Work 56

6.1 Comparison of Psychophysiological Features Between Devices

6.1.1 Comparison Set-up

As mentioned in Section 3.1.4, the signals recorded from the stationary devices are ECG, part of EDA(SCL), and fEMG (from COR and ZYG muscles) in sampling frequencies of 1kHz, 250Hz and 1kHz respectively. Regarding the wearable devices, the signals recorded using the band are PPG (1 Hz) and EDA (0.2 Hz) and using the ring only EDA (3 Hz) was recorded. The signals that can be compared are ECG between the stationary device and the band, SCL between all devices, and EDA between the two wearable devices.

In a previous study [60], the same data (but without data multiplication) were used and the correlation between the ECG and SCL signals of the stationary device and the band was analysed. The results shown strong correlation between the ECG features and weak correlation between SCL features. In this study the same signals are compared using the dataset generated by data multiplication in the 10-second window, in order to find out if the findings remain the same in the window-size with the best performance.

Moreover, the remaining signals are compared, which are EDA of the band and the ring and SCL of the stationary device and the ring. The data used to perform these comparisons, are the data generated by data multiplication in the 40-second window. For each signal, feature extraction was held as explained in Section 4. The reason why only the data from the 40-second window were used for these comparisons is because only in these data it was possible to extract all features from all devices. In other window sizes, no healthy SCR peaks were observed in ring's EDA signal, hence SCR-related features could not be extracted from the ring, due to errors in signal recording (instant EDA was recorded instead of raw EDA). This comes in contrast to the previous analysis, which used 10-second intervals.

6.1.2 Results

Pearson's correlation coefficient and two-tailed p-value were computed. Additionally, as Pearson's correlation may be misleading (i.e., high correlation does not always imply high agreement), the Bland-Atman scatter plots [61] were created.

Bland-Atman scatter plot is an effective representation of the relationship between two paired variables using the same scale. The horizontal axis represents the mean values of two measurements and the vertical axis the difference between the two measurements. The plot includes three reference lines. The first one is the upper limit of agreement, which is equal to $AV + 1.96 \times SD$, where AV is the average value of the measurements and SD is their Standard Deviation. The second one is the mean difference between the measurements and the last one is the lower limit of agreement, that is equal to $AV - 1.96 \times SD$. The upper and lower limits of agreement are represented as dashed lines, while the mean difference is a solid line. If the mean difference is close to zero, then the two measurements agree perfectly. Also, if the values are between the dashed lines, then the measurements agree.

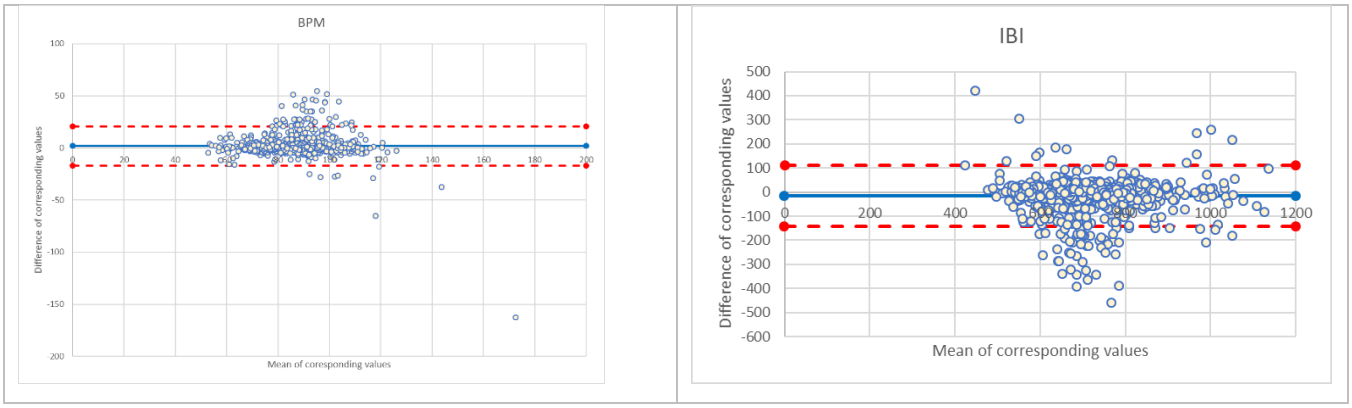
A Bland-Atman scatter plot was constructed for each feature that was being compared.

6.1.2.1 Results Comparing ECG and SCL Signals of BIOPAC and Microsoft Band 2

Correlations between devices were examined on the mean of each feature in the 10-second intervals. The findings of the comparison between ECG and SCL features of BIOPAC and Microsoft Band 2, are shown in Table 6.3 and Table 6.2.

Feature	Mean in BIOPAC	Mean in Microsoft Band 2	r-value (Correlation)	p-value
IBI	699.00	719.00	0.02	0.48
BMP	87.60	85.30	0.04	0.15
SDNN	48.00	39.50	0.09	0.00
SDSD	29.50	28.70	0.07	0.02
RMSSD	48.40	58.60	0.06	0.03
pNN20	0.51	0.52	-0.01	0.71
pNN50	0.19	0.26	0.01	0.60
HRMAD	31.50	25.50	0.06	0.05
Mean SCL	48.04	0.61	0.01	0.66
Standard Deviation SCL	0.09	0.01	0.02	0.43

Table 6.1 Results of comparison between ECG and SCL features of BIOPAC and Microsoft Band 2



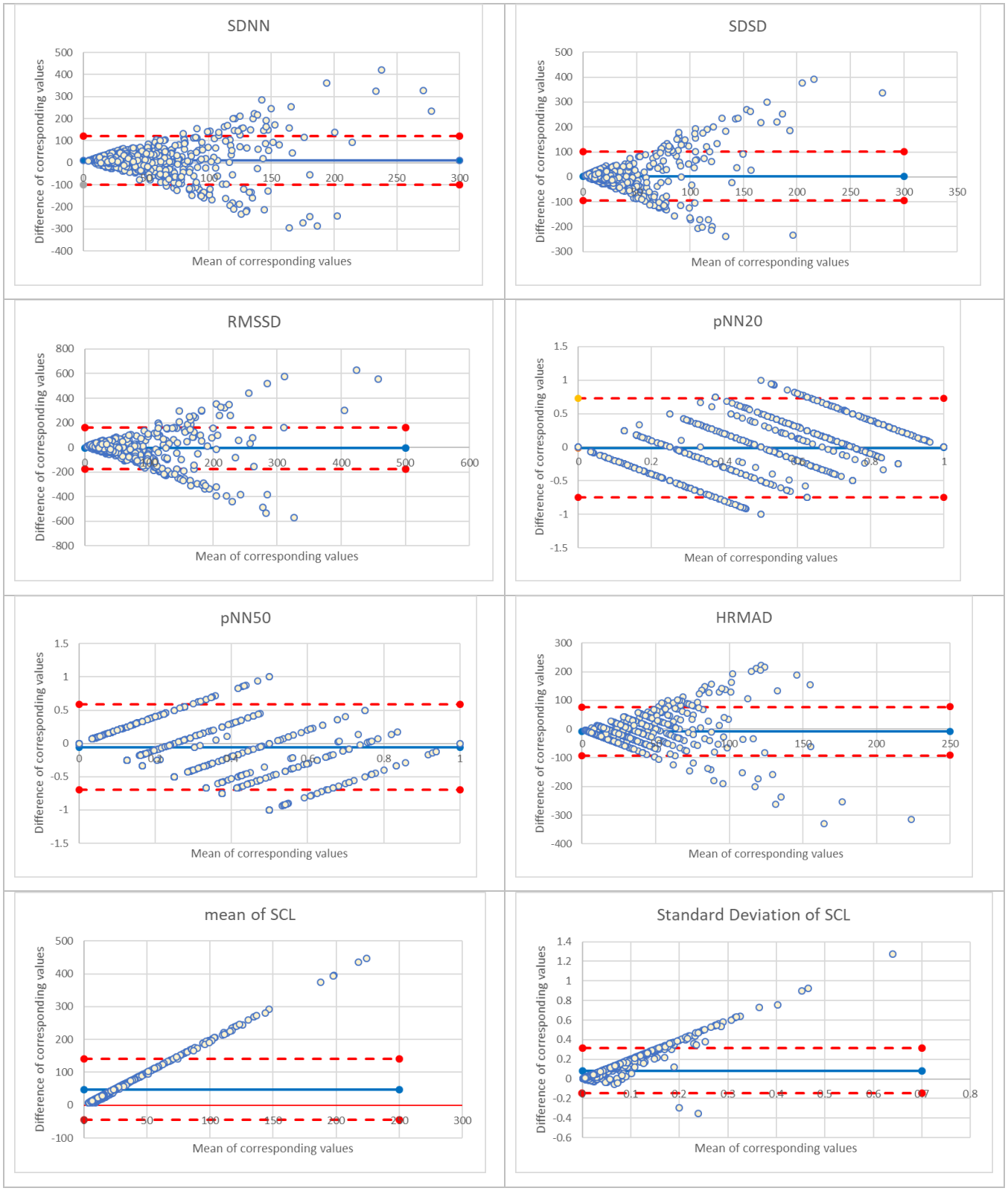


Table 6.2 Scatter plots regarding comparison between ECG and SCL features of BIOPAC and Microsoft Band 2

In some of the features (SDNN, SDSD, RMSSD, and HRMAD) the p-value is less than or equal to 0.05 (indicating that the correlation is statistically significant), and the correlation (r-value)

between the two devices is weak ($r < 0.40$). In the rest features the p-value is higher than the significance level ($\alpha = 0.05$). Thus, we fail to reject the null hypothesis and we conclude that their correlation is not statically significant.

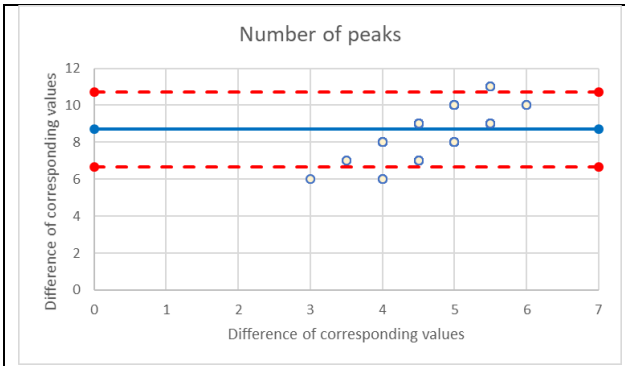
Looking at the scatter plots, we can observe that features that come from the ECG signal agree, while the two features that are related to SCL (mean value and standard deviation) are not in agreement. Thus, the findings remain the same as in the previous study [35], when considering this method.

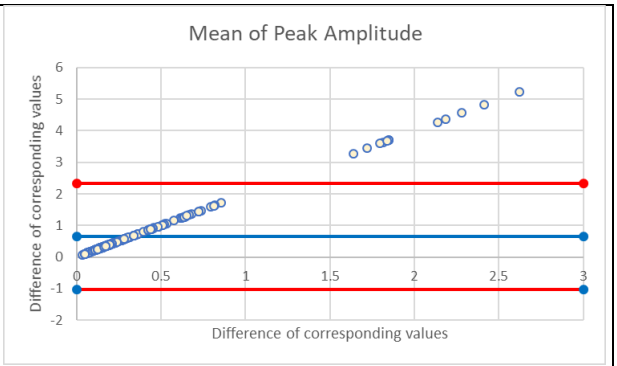
6.1.2.2 Results Comparing EDA Signal of Microsoft Band 2 and Moodmetric Smart Ring

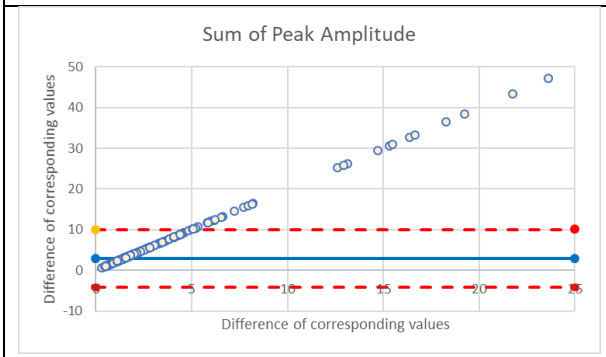
The correlation between the devices was examined on the mean of each feature in the 40-second intervals. The findings are shown in Table 6.3Table 6.3 and Table 6.4.

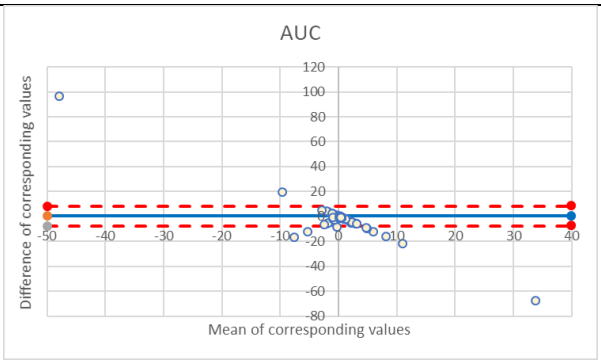
Feature	Mean in Microsoft Band 2	Mean in Moodmetric Smart Ring	r-value (Correlation)	p-value
peak count	9.13	$4.29E \times 10^{-1}$	9.17×10^{-2}	0.16
mean peak amplitude	0.64	2.63×10^{-5}	9.40×10^{-3}	0.89
sum peak amplitude	5.73	2.63×10^{-5}	9.78×10^{-3}	0.88
AUC	-0.15	9.17×10^{-1}	-2.25×10^{-3}	0.97
mean EDA	0.64	6.12×10^{-5}	6.59×10^{-2}	0.31
Standard Deviation EDA	0.01	3.07×10^{-7}	-8.73×10^{-2}	0.18
kurtosis EDA	0.11	-3.23×10^{-2}	-1.35×10^{-2}	0.84
skew EDA	0.180	-1.32×10^{-1}	8.33×10^{-2}	0.2
mean SCR	0.01	2.42×10^{-7}	1.71×10^{-2}	0.79
Standard Deviation SCR	0.01	1.15×10^{-7}	-4.57×10^{-2}	0.48
Mean SCL	0.63	6.10×10^{-5}	2.58×10^{-2}	0.69
Standard Deviation SCL	0.01	1.75×10^{-7}	-7.17×10^{-2}	0.27

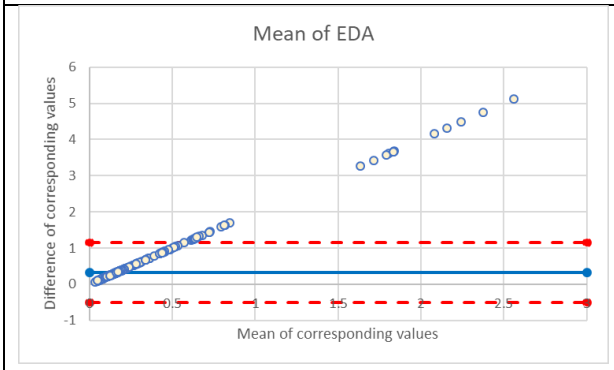
Table 6.3 Results of comparison between EDA features of Microsoft Band 2 and Moodmetric Smart Ring

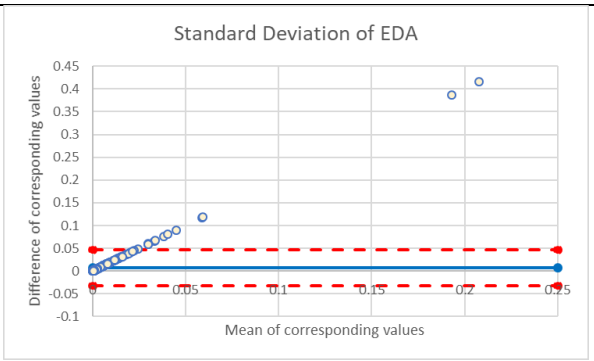

Number of peaks
Difference of corresponding values
Difference of corresponding values

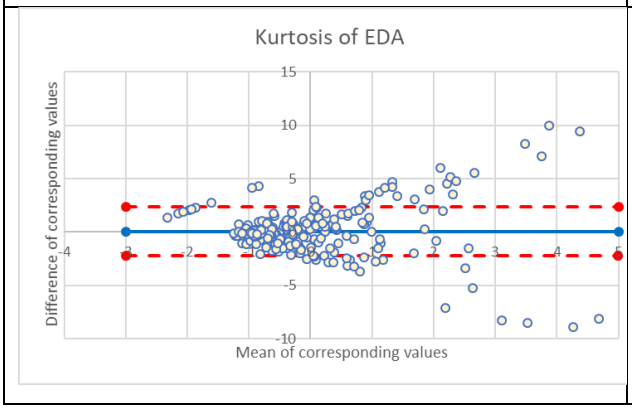

Mean of Peak Amplitude
Difference of corresponding values
Difference of corresponding values

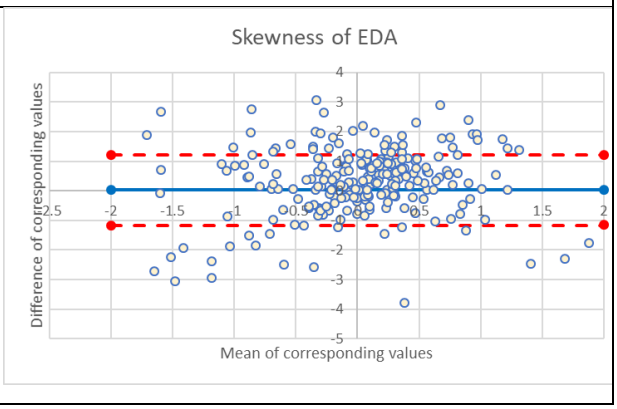

Sum of Peak Amplitude
Difference of corresponding values
Difference of corresponding values


AUC
Difference of corresponding values
Mean of corresponding values


Mean of EDA
Difference of corresponding values
Mean of corresponding values


Standard Deviation of EDA
Difference of corresponding values
Mean of corresponding values


Kurtosis of EDA
Difference of corresponding values
Mean of corresponding values


Skewness of EDA
Difference of corresponding values
Mean of corresponding values

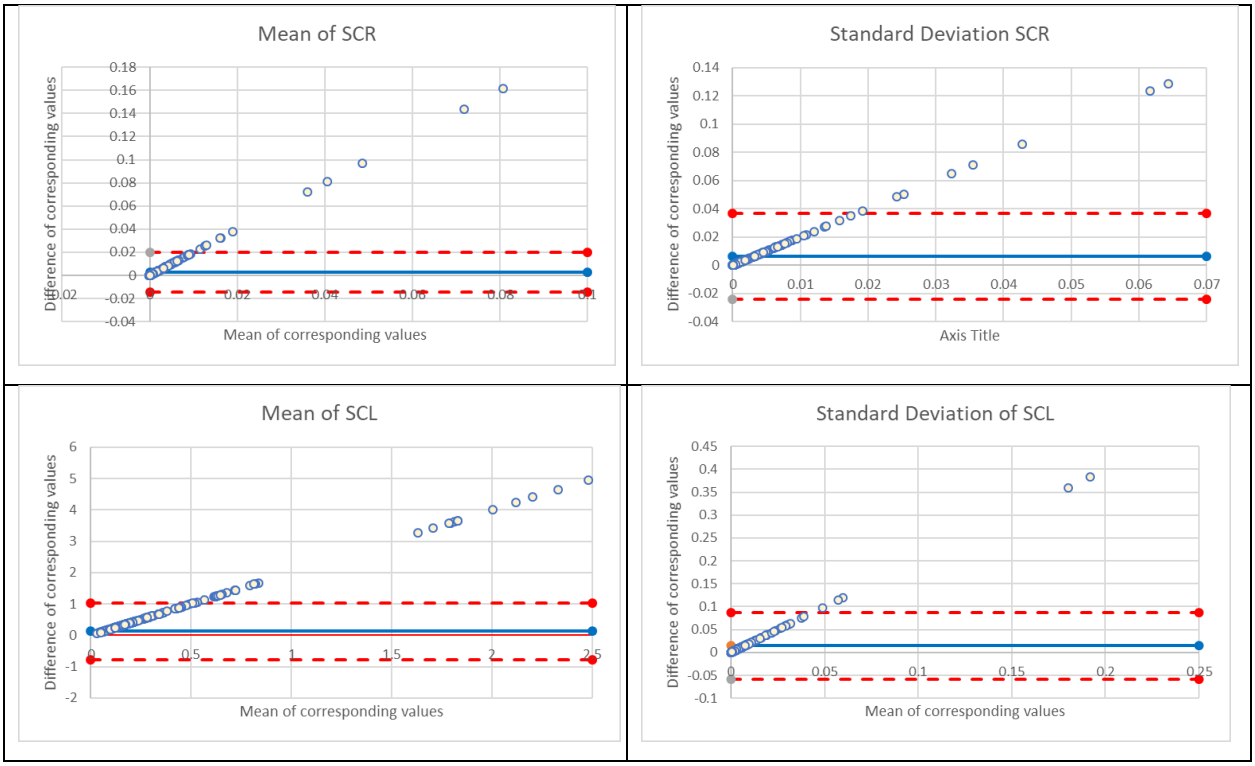


Table 6.4 Scatter plots regarding comparison between EDA features between Microsoft Band 2 and Moodmetric Smart Ring

The p-value is higher than the significance level ($\alpha = 0.05$) in all features. Hence, we fail to reject the null hypothesis and, based on Pearson's correlation, we conclude that the correlation is not statically significant. Additionally, the mean values of each feature are not close to each other when measured with the two devices. However, in Bland-Altman scatter plots a large number of points are between the two dashed lines (limits of agreement), indicating that the measurements agree.

6.1.2.3 Results Comparing SCL Signal of BIOPAC and Moodmetric Smart Ring

The whole EDA signal of the stationary device was not recorded, thus only SCL could be compared between the stationary device and the ring. The results are shown in Table 6.5 and Table 6.6.

Feature	Mean in BIOPAC	Mean in Moodmetric Smart Ring	r-value (Correlation)	p-value
Mean SCL	8.93	6.10×10^{-5}	0.15	0.02

Feature	Mean in BIOPAC	Mean in Moodmetric Smart Ring	r-value (Correlation)	p-value
Standard Deviation SCL	0.19	1.75×10^{-7}	0.33	1.29×10^{-7}

Table 6.5 Results of comparison between SCL features of BIOPAC and Moodmetric Smart ring

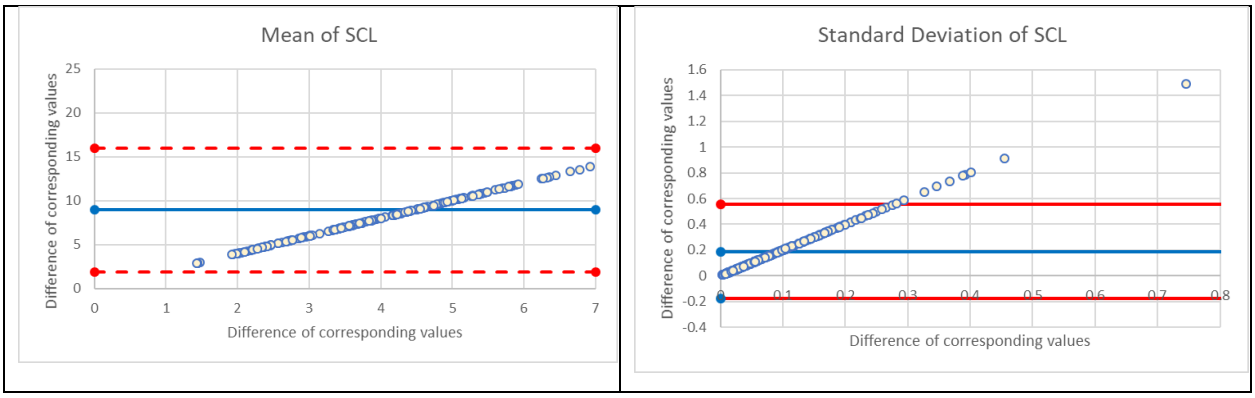


Table 6.6 Scatter plots regarding comparison between SCL features between BIOPAC and Moodmetric Smart Ring

As shown in the table, the p-value is lower than 0.05. Thus, the result is statistically important. The correlation between the two devices is weak in both features (r <0.40). Looking at the scatter plots, it can be noticed that values remain between the two dashed lines, which implies that SCL measurements from the ring and the stationary device agree, despite the fact that their mean values are not close.

6.2 Comparison of Results with Previous Work

This section compares the results of the current work with the three previous related analyses that were described in Section 3.2. Table 6.7 shows the candidate features and the selected features in each analysis, as well as the selected classifier in each one.

	Selected Features	Best Classifier Overall	Average Accuracy of the Best Classifier (%)
Galazis, 2017 [1]	Smoking: GSR	RF	66
	Eating Disorder: ECG, BI-AAQ (questionnaire)		
	Anxiety: COR		
Trigeorgi, 2018 [18]	Smoking: ECG_rmssd	RF	73
	Eating Disorder: ECG_sdnn, COR_mean		
	Anxiety: ECG_bpm, ECG_rmssd, ECG_pnn20, COR_mean		
Demosthenous, 2019 [32]	All: GSR_mean, COR_mean	BDT (RWM)	90
Current study	Pain: SCRwatch_meanPeakAmp, ECG_heartrate	BDT (RWM)	85

Table 6.7 Results of previous analyses and current work

As it can be observed, there is no certain family of features or combination that yielded the best performance in all cases. In the three previous studies, the most effective features appear to come from ECG and COR signals, while GSR is not that popular. In the current study, SCR peak amplitude, which was extracted from GSR, belongs to the combination that gave the best results. Thus, if more features were extracted from all the available signals in previous studies, maybe GSR would occur more often in the selected features.

In terms of the most effective Machine Learning Algorithm, Random Forest performed best in the two earliest studies, while Bagging Decision Tree was the best in the last and current work. The accuracy of the most effective classifier is much lower in the two earliest works (66% and 77% respectively), whereas BDT had a very good performance in the last two works (90 and 85% respectively). This may not be due to the algorithm, but due to the fact that the first two works did not used Rectangular Window Methodology to multiply the data and relied on the hypothesis that the class of each individual remained the same throughout the experiment. Lastly, in the analysis

of 2019 a higher accuracy was succeeded. A possible reason is because in that study the data used for the classification were recorded using BIOPAC, which is more reliable and precise than Microsoft Band, which has limitations due to sensors, power and memory. Nevertheless, the accuracy achieved is very satisfactory.

Chapter 7

Discussion

7.1 Summary 59

7.2 Future Work 57

7.1 Summary

This thesis is the continuation of a series of experiments and analyses of emotional coping using psychophysiological features. The data were acquired from an experiment conducted by the Department of Psychology of the University of Cyprus. The experiment is known as the Cold Pressor Task, and involves placing a hand in an ice-cold water and is used to examine pain threshold and tolerance. During the experiment, three monitoring devices were used; BIOPAC, Microsoft Band 2, and Moodmetric Smart Ring. The signals recorded from the stationary devices are ECG, SCL, and fEMG (from COR and ZYG muscles) in sampling frequencies of 1kHz, 250Hz and 1kHz respectively. From the band the signals recorded are PPG and EDA in sampling frequencies of 1Hz and 0.2 Hz respectively. Using the ring only EDA (3 Hz) was recorded.

Due to the fact that the dataset was relatively small (80 participants), a number of artificial samples were produced using the existing data. To do this Rectangular Window Methodology was utilized, which starts from the beginning of each signal and takes a chunk of the data and treats is as an individual data sample. This procedure was repeated for four different window sizes (10, 20, 30, and 40 seconds) and four datasets were generated. The work described from this point onward, was done in all four datasets. From each signal, multiple features were extracted. Regarding ECG and HRV signals, main focus was given in time-domain measures, which include finding peaks in

the signal and extracting information from them, like the time passed between each signal, etc. The EDA signal was split into its two components; SCL and SCR. Statistical metrics were extracted from SCL, while a similar procedure to the one followed for ECG was done in SCR. Statistical features were extracted from COR and ZYG muscles signals.

The topic that was examined the most in this thesis, is the selection of the most relevant features. Three different feature selection methods were analyzed; Wrapper, Embedded, and Filter Methods. From each method, one technique was selected. The first technique, which belongs to Wrapper Methods, is Exhaustive Feature Selection, which includes training a model using all possible subsets of features and selecting the one with the best score. Because there were a large number of possible features, only combinations of size three or less were examined. Feature Importance, which belongs to Embedded Methods, ranks all the features from the most relative to the least. The results of these two techniques were similar. In stationary devices, SCL and statistical properties of COR and ZYG signals yielded good results, while in wearable devices, the average value of EDA signal, SCR amplitude, and heart rate had the best performance. The technique chosen from Filter Methods is Correlation Coefficient, which computes the Pearson correlation between all possible pairs of features aiming to exclude from the selection the features that have high correlation. This technique showed that none of the features are directly related to the target variable and that only features coming from the same signal have a strong correlation among them. Putting it all together, it was observed that features coming from the band had quite good results, so both of the features that were finally chosen came from this device. The subset selected is the mean value of heart rate and the mean value of SCR amplitude.

Five classifiers were trained and compared. These are Adaptive Boosting, Gradient Boosting Decision Tree, Random Forest, and Extra Trees. All of these classifiers use decision trees to fit the model. To selected the best number of trees for each algorithm, the data from the 20-second window were used and five numbers of trees were tested: 25, 50, 100, 200 and 300. BDT performed better with 50 estimators, and the rest achieved the best performance when having 100 estimators. Thereafter, all algorithms were trained on all four datasets. The classifier that showed the best performance is Bagging Decision Tree, with an excellent average accuracy of 85% and an average standard deviation of 0.03%, demonstrating that wearable devices can perform really well in classifying the individual to the two pain-coping categories.

Moreover, the signals that the monitoring devices had in common were compared. The findings are different when using different methods to compare the data. Pearson correlation shows no correlation between most of the features, while Bland-Altman scatter plots show agreement in the measurements in most of the cases.

Furthermore, when comparing the results of the current study with the previous studies it was observed that data multiplication using RWM improves the performance of the classifiers. Moreover, data acquired by the band can give results with performance close to that acquired by the stationary device.

7.2 Future Work

Great results were yielded from the available data. However, there are some changes that could be done in data collection that may bring even better results in related studies. Firstly, it was observed that the monitoring devices lacked of synchronization and this may have affected the comparison of the devices. Additionally, the EDA signal that was recorded from the band, included several zeros in multiple data samples. There are two possible reasons: either the device used was of low quality or the sampling rate in which it was set to record was not sufficient. Although the band used could reach a sampling rate of 5Hz, it was set to 0.2 Hz to save memory and power.

Moreover, in future work, the whole EDA signal could be extracted from the stationary devices as well (not only its tonic level), in order to compare it with the whole EDA signal of the wearable devices. This would be an indicator of whether it is admissible to use features extracted from SCR to train the classifiers.

Lastly, the two data samples were unbalanced. More precisely, there were a lot more data samples that belonged to the ‘functional’ category, rather than the ‘dysfunctional’ one. Although a number of methodologies were used to tackle this issue, it could have resulted in classifiers getting biased, and decreasing their performance.

References

[1] C. Galazis, "Non-Intrusive Physiological Wearable Devices for Identifying Individual Difference Parameters Using Supervised Classification Learning Algorithms," Computer Science Department, Univeristy of Cyprus, Nicosia, Cyprus, 2017.

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

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

[4] C. E. Ackerman, "How Does Acceptance And Commitment Therapy (ACT) Work?," 29 March 2022. [Online]. Available: https://positivepsychology.com/act-acceptance-and-commitment-therapy/. [Accessed 23 May 2022].

[5] D. C. Swanson, Signal Processing for Intelligent Sensor Systems with MATLAB, Boca Raton: CRC Press, 2011.

[6] Y. Sattar and L. Chhabra, in *Electrocardiogram,* StatPearls Publishing, 2021.

[7] E. Ashley and J. Niebauer, in *Cardiology Explained*, London, Remedica, 2004.

[8] B. Farnsworth, "What Is ECG and How Does It Work?," August 3 2021. [Online]. Available: https://imotions.com/blog/what-is-ecg/. [Accessed 7 May 2022].

[9] D. Castaneda, A. Esparza, M. Ghamari, C. Soltanpur and H. Nazeran, "A review on wearable photoplethysmography sensors and their potential future applications in health care," *Int J Biosens Bioelectron,* vol. 4, no. 4, pp. 195-202, 2018.

[10] A. Alqaraawi, A. Alwosheel and A. Alasaad, "Heart rate variability estimation in photoplethysmography signals using Bayesian learning approach," *Healthc Technol Lett.*, vol. 3, no. 2, pp. 136-142, 2016.

[11] M. Benedek and C. Kaernbach, "A continuous measure of phasic electrodermal activity," *Journal of Neuroscience Methods,* vol. 190, no. 1, pp. 80-91, 2010.

[12] "Galvanic Skin Response (GSR): The Complete Pocket Guide," iMotions, 2020.

[13] J. Wilson, "What Is Facial EMG and How Does It Work?," iMotions, 2018.

[14] T. Marur, Y. Tuna and S. Demirci, "Facial anatomy," *Clinics in Dermatology,* vol. 32, no. 1, pp. 14-23, 2014.

[15] J. Tong, M. J. Lopez and B. C. Patel, "Anatomy, Head and Neck, Eye Orbicularis Oculi Muscle," StatPearls Publishing, Treasure Island (FL), 2021.

[16] J. T. Cacioppo, R. E. Petty, M. E. Losch and H. S. Kim, "Electromyographic Activity Over Facial Muscle Regions Can Differentiate the Valence and Intensity of Affective Reactions," *Journal of Personality and Social Psychology,* vol. 50, no. 2, pp. 260-268, 1986.

[17] R. Brooks and K. Dahlke, "Understanding the 3 Categories of Machine Learning – AI vs. Machine Learning vs. Data Mining 101 (part 2)," 17 October 2017. [Online]. Available: https://www.guavus.com/ai-vs-machine-learing-vs-data-mining-whats-big-difference-part-2/. [Accessed 23 May 2022].

[18] Ά. Τριγιώργη, «Εξόρυξη Γνώσης από Ψυχοφυσιολογικά Δεδομένα και Συγκριτική Αξιολόγηση Αλγορίθμων Μηχανικής Μάθησης,» 2018.

[19] S. Besetty, "Decision Tree," 22 June 2021. [Online]. Available: https://www.geeksforgeeks.org/decision-tree/. [Accessed 16 April 2022].

[20] sunil, "Quick Introduction to Boosting Algorithms in Machine Learning," 9 November 2015. [Online]. Available: https://www.analyticsvidhya.com/blog/2015/11/quick-introduction-boosting-algorithms-machine-learning/. [Accessed 16 April 2022].

[21] V. Kurama, "Gradient Boosting In Classification: Not a Black Box Anymore!," 2020. [Online]. Available: https://blog.paperspace.com/gradient-boosting-for-classification/. [Accessed 18 April 2022].

[22] J. Brownlee, "A Gentle Introduction to Ensemble Learning Algorithms," 19 April 2021. [Online]. Available: https://machinelearningmastery.com/tour-of-ensemble-learning-algorithms/. [Accessed 16 April 2022].

[23] S. E. R, "Understanding Random Forest," 17 June 2021. [Online]. Available: https://www.analyticsvidhya.com/blog/2021/06/understanding-random-forest/. [Accessed 16 April 2022].

[24] P. Aznar, "What is the difference between Extra Trees and Random Forest?," 17 June 2020. [Online]. Available: https://quantdare.com/what-is-the-difference-between-extra-trees-and-random-forest/#:~:text=Random%20Forest%20chooses%20the%20optimum,randomization%20but%20still%20has%20optimization.. [Accessed 18 April 2022].

[25] J. Brownlee, "A Gentle Introduction to k-fold Cross-Validation," 23 May 2018. [Online]. Available: https://machinelearningmastery.com/k-fold-cross-validation/. [Accessed 21 April 2022].

[26] R. Toshniwal, "How to select Performance Metrics for Classification Models," 9 January 2020. [Online]. Available: https://medium.com/analytics-vidhya/how-to-select-performance-metrics-for-classification-models-c847fe6b1ea3. [Accessed 8 April 2022].

[27] [Online]. Available: https://www.biopac.com/wp-content/uploads/MP150-Systems.pdf. [Accessed 22 May 2022].

[28] "Hardware Specifications BIOPAC MP150," [Online]. Available: https://imotions.com/hardware/biopac-mp150/. [Accessed 23 May 2022].

[29] "ACQKNOWLEDGE 3.9 SOFTWARE GUIDE," [Online]. Available: https://www.biopac.com/manual/acqknowledge-3-9-software-guide/. [Accessed 22 May 2022].

[30] D. S. McConnell, N. S. Chudy and J. Smither, "The Performance of Microsoft Band 2: A Validity and Reliability Study," in *Human Factors and Applied Psychology Student Conference*, Daytona, 2016.

[31] [Online]. Available: https://moodmetric.com/. [Accessed 22 May 2022].

[32] G. Demosthenous, "Machine Learning Approach to Predict Emotional Coping Using Psychophysiological Signals," Master Thesis, Nicosia, 2019.

[33] C. S. Hayes, B. J. Luoma, W. F. Bond, A. Masuda and J. Lillis, "Acceptance and Commitment Therapy: Model, processes and outcomes," *Behaviour Research and Therapy*, vol. 44, no. 1, pp. 1-25, 2006.

[34] E. K. Sandoz, K. G. Wilson, R. M. Merwin and K. K. Kellum, "Assessment of body image flexibility: The Body Image-Acceptance and Action Questionnaire," *Journal of Contextual Behavioral Science,* vol. 2, no. 1-2, pp. 39-48, 2013.

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

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

[37] S. Sanamdikar, S. Hamde and V. Asutkar, "Analysis and classification of cardiac arrhythmia based on general sparsed neural network of ECG signals," *SN Applied Sciences,* vol. 2, no. 7, 2020.

[38] E. Morgan, "All About HRV Part 2: Interbeat Intervals and Time Domain Stats," MindWare, Greenwich, 2017.

[39] B. Farnsworth, "Heart Rate Variability – How to Analyze ECG Data," 19 July 2019. [Online]. Available: https://imotions.com/blog/heart-rate-variability/. [Accessed 22 April 2022].

[40] F. Shaffer and J. P. Ginsberg, "An Overview of Heart Rate Variability Metrics and Norms," Front Public Health, 2017.

[41] P. v. Gent, "Analyzing a Discrete Heart Rate Signal Using Python – Part 2," 21 March 2016. [Online]. Available: http://www.paulvangent.com/2016/03/21/analyzing-a-discrete-heart-rate-signal-using-python-part-2/. [Accessed 3 May 2022].

[42] 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,* vol. 12, no. 4, pp. 857-869, 2021.

[43] D. Makowski, T. Pham, Z. J. Lau, J. C. Brammer, F. Lespinasse, H. Pham, C. Schölzel and S. H. A. Chen, "NeuroKit2: A Python toolbox for neurophysiological signal processing," *Behavior Research Methods,* vol. 4, no. 1689-1696, p. 53, 2021.

[44] M. Memar and A. Mokaribolhassan, "Stress level classification using statistical analysis of skin conductance signal while driving," *SN Applied Sciences,* vol. 3, no. 1, 2021.

[45] B. Farnsworth, "Skin Conductance Response – What it is and How to Measure it," 29 January 2019. [Online]. Available: https://imotions.com/blog/skin-conductance-response/. [Accessed 22 April 2022].

[46] W. Boucsein, Electrodermal Activity, New York: Springer, 2012.

[47] H. F. Posada-Quintero and K. H. Chon, "Innovations in Electrodermal Activity Data Collection and Signal Processing: A Systematic Review," *Sensors,* vol. 20, no. 2, p. 479, 2020.

[48] D. Thulkar, T. Bhaskarwar and S. T. Hamde, "Facial electromyography for characterization of emotions using LabVIEW," *2015 International Conference on Industrial Instrumentation and Control (ICIC),* pp. 683-686, 2015.

[49] A. Gupta, "Feature Selection Techniques in Machine Learning," 10 October 2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/. [Accessed 12 April 2022].

[50] H. Liu, M. Zhou and Q. Liu, "An Embedded Feature Selection Method for," *IEEE/CAA JOURNAL OF AUTOMATICA SINICA,* vol. 6, no. 3, pp. 703-715, 2019.

[51] V. Verma, "A comprehensive guide to Feature Selection using Wrapper methods in Python," 24 October 2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/10/a-comprehensive-guide-to-feature-selection-using-wrapper-methods-in-python. [Accessed 12 April 2022].

[52] E. Scornet, "Trees, forests, and impurity-based variable importance in regression," 2020.

[53] C. Lee, "Feature Importance Measures for Tree Models — Part I," 28 October 2017. [Online]. Available: https://medium.com/the-artificial-impostor/feature-importance-measures-for-tree-models-part-i-47f187c1a2c3. [Accessed 13 April 2022].

[54] A. Dubey, "Feature Selection Using Random forest," 15 December 2018. [Online]. Available: https://towardsdatascience.com/feature-selection-using-random-forest-26d7b747597f. [Accessed 13 April 2022].

[55] S. Yemulwar, "Feature Selection Techniques," 27 September 2019. [Online]. Available: https://medium.com/analytics-vidhya/feature-selection-techniques-2614b3b7efcd. [Accessed 5 April 2022].

[56] A. Gupta, "Feature Selection Techniques in Machine Learning," 10 October 2020. [Online]. Available: https://www.analyticsvidhya.com/blog/2020/10/feature-selection-techniques-in-machine-learning/. [Accessed 15 April 2022].

[57] G. Brown, A. Pocock, M.-J. Zhao and M. Lujan, "Conditional Likelihood Maximisation: A Unifying Framework for Information Theoretic Feature Selection," *Journal of Machine Learning Research,* vol. 13, no. 1, pp. 27-66, 2012.

[58] "Pearson correlation coefficient," 19 May 2022. [Online]. Available: https://en.wikipedia.org/wiki/Pearson_correlation_coefficient. [Accessed 20 May 2022].

[59] M. L. Waskom, "seaborn: statistical data visualization," *Journal of Open Source Software,* vol. 6, no. 60, p. 3021, 2021.

[60] K. Pinelopi, T. Andria, G. Chryssis, G. A. T, P. Georgia and K. Maria, "Comparing apples and oranges using wearable vs. stationary devices to analyze psychophysiological data," University of Cyprus, Nicosia, 2021.

[61] J. M. Bland and D. Altman, "Statistical methods for assessing agreement between two methods of clinical measurement," *Lancet,* vol. 1, no. 8476, pp. 307-310, 1986.