htw saar

Systems Neuroscience
& Neurotechnology Unit

Neuroergonomic Assessment of Human Robot Interaction

Master Thesis

Systems Neuroscience & Neurotechnology Unit
Saarland University of Applied Sciences
Faculty of Engineering

Submitted by	:	Dominik Limbach, B.Sc.
Matriculation Number	:	3662306
Course of Study	:	Biomedical Engineering (Master)
Specialisation	:	Neural Engineering
First Supervisor	:	Prof. Dr. Dr. Daniel J. Strauss
Second Supervisor	:	Dr. Lars Haab

Saarbrücken, January 13, 2020

Abstract

In recent years a growing body of research on Human Robot Interaction has reported the benefits of signal based emotion recognition for both system performance and the mental health of the user.

However, the majority of studies employed devices and measures that are obtrusive in nature and therefore far too impractical for the application in a realistic setting.

The present thesis introduces a novel approach for unobtrusive emotion recognition in Human-Robot-Interaction. Using the Empatica E4, a medical-grade wearable, we designed a wireless monitoring system that is capable of performing feature based emotion classification. In order to build the system we conducted a data collection experiment, recording three physiological signals (blood volume pulse, galvanic skin response, and skin temperature) of 14 individuals. During the experiment the participants had to first perform two cognitive tasks (Mental Arithmetic Stress Test, STROOP Test) and then undergo two sessions of visual stimulation. Each session used a set pictures specifically selected (from the International Affective Picture System) to elicit a certain emotional state. Afterwards, we handled feature extraction and created a variety of data-, and feature sets for the purpose of machine learning. Lastly, we evaluated the performance of four machine learning algorithms (k-Nearest Neighbor, Support Vector Machines, Random Forest Classifier, and Neural Networks) based on their classification accuracy. With this approach we were able to classify a total of 6 different emotional states, with accuracies as high as 58% for individual classifiers, and an average of 44% for all machine learning algorithms. Further, we were able to achieve even higher performances for binary classification tasks, aiming to identify specific emotional states. Considering the average performances of all four algorithms we achieved accuracies of 89% for stress sessions, 72% for visual stimulation sessions, and 91% for resting intervals.

Closing statement

Zusammenfassung

In den letzten Jahren sind Wearables zu einer festen Größe in unserer Gesellschaft geworden. In der Form von Smartwatches und Fitness-Armbändern haben Wearables Einzug in annähernd dreißig Prozent aller deutschen Haushalte gehalten. Neben den Grundfunktionen eines Telefons oder einer Uhr bieten diese praktischen Geräte eine Bandbreite an Funktionen. Viele dieser Zusatzfunktionen befassen sich mit dem Messen der Vitalparameter des Trägers zur Früherkennung von Krankheiten oder zur Optimierung der sportlichen Leistungsfähigkeit. Forschungsarbeiten der letzten Jahre lassen das Potential der Anwendung von psychophysiologischem Monitoring in besonders stressvollen Arbeitsplätzen erahnen, ein Gebiet in dem Wearables nur selten anzutreffen sind. Aus diesem Grund haben wir ein System für Arbeiter an besonders stressvollen Arbeitsplätzen, insbesondere collaborative Arbeitsplätze mit Mensch-Roboter-Interaktion (HRI). Das System basiert auf einem Wearable, welches am Handgelenk getragen wird, und ist in der Lage den mentalen Zustand des Trägers zu beurteilen. Durch den Einsatz eines solchen Systems sind wir in der Lage neuroergonomische Atbeitsplätze zu schaffen die sich an die cognitive Leistungsfähigkeit eines Arbeiters anpassen und somit den Einfluss von Stress, als einen der führenden Gründe für Krankheiten und Verletzungen unter der arbeitenden Bevölkerung, drastisch zu senken.

Acknowledgements

First of all, I want to give thanks to Prof. Dr. Dr. Daniel J. Strauss for providing me with such an interesting topic and granting me the freedom of working in such an independent manner. This allowed me to fully invest myself in the design and realization of this project.

Further, I would like to thank the entire staff of the Systems Neuroscience and Neurotechnology Unity for providing such a kind and supportive environment. In particular, I want to say thanks to Dr. Lars Haab, who not only was always content to help but also willing to supervise me on yet another one of my projects.

Last but not least I want to express my deepest gratitude to my family and friends. Your continuous support throughout the years is what kept me motivated and made this thesis possible in the first place.

Thank You!

Declaration

I hereby declare that I have authored this work independently, that I have not used other than the declared sources and resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources. This work has neither been submitted to any audit institution nor been published in its current form.

Saarbrücken, January 13, 2020

Dominik Limbach, B.Sc.

Contents

1 Introduction

In this chapter we will introduce the role of neuroergonomics in adaptive automation work environments and look at some of the problems that led to the current state of the field. Additionally, we will provide background information on crucial topics, such as psychophysiological measures, machine learning techniques, and emotion classification.

1.1 Introduction to the Field

Neuroergonomics is often described as the study of brain and behavior at work. As the name suggests, neuroergonomics is comprised of two disciplines, neuroscience and ergonomics (also known as human factors). Neuroscience is concerned with the structure and the function of the brain. It is a highly interdisciplinary body of research spanning disciplines such as physiology, psychology, medicine, computer science, or mathematics. Human factors on the other hand is focused on examining the human use technology at work or other real-world settings. As the intersection of these two fields, neuroergonomics addresses both the brain and humans at work, but even more so their dynamic interaction [1] . By understanding the neural bases of perceptual and cognitive functions, such as seeing, hearing, planing and decision making in relation to technology and settings in the real world, neuroergonomics strives to develop new optimization methods for various areas of applications. The additional value neuroergonomics can provide, compared to 'traditional' neuroscience and 'conventional' ergonomics, promises substantial economical benefits as well as significant improvements to health care and therefore society at large. In the scope of this thesis, we will focus primarily on applications in work settings such as modern automated systems. An area in which the effects of neuroergonomics are expected to be even greater, considering the difficulty of obtaining measures of overt behavior [1].
Automated systems, in one form or another have been present in almost every branch of industry for the better part of what has been two centuries. Automation itself can be defined as the creation and application of technology by which the production and delivery of various goods and services is controlled and monitored with minimal human assistance. Automation can be encountered in many different places, from a simple control loop in a hydraulic system up to an artificial intelligence handling emergency breaking in autonomic cars.Static automation is the original form of automation. In this form, automation is an all-or-none technology either performing a task for us or

not [2]. Although, there are numerous benefits (i.e. relieving workers from the strain of performing repetitive tasks) to it, there also is increasing evidence that static automation comes at the price of impaired decision making, manual skill degradation, loss of situational awareness, and monitoring inefficiency [2]. These problems stem from the significant change of roles that automation causes in a work environment. What this means is that workers, once active operators of machines and technology, are now passive monitors who often face monitoring workloads that are inherently different and often significantly higher when compared to the manual control conditions of a non automated work environment. The continuous exposure to workloads that are either too high or too low can have dramatic effects on human-system performance, thereby potentially compromising safety [3]. This has already been indicated by the work of Parasuraman et al. (1993, 1994) who tested subjects in a multi-task flight simulator with optionally automatable components and reported a substantial decrease of human operator detection of automation failures after short periods of time in a static automation scenario with constant task assignment of both operator and the automated system [2]. However, we must not forget the consequences these highly automated and ever so stressful workplaces have on human operators specifically. Studies by Cooper et al. showed that working in a stressful environment increases the risk of suffering physical illness or symptoms of psychological distress, as well as work related accidents and injuries [4].
We will now take a look at what is called adaptive automation. Adaptive automation has been introduced to resolve the abovementioned issues of statically automated systems. Whereas static automation is considered to be an agent working for the operator, adaptive automation is viewed as an interactive aid working with the operator [2]. It attempts to optimize system performance by adjusting the task assignment between the human operator and automation dynamically. This task reallocation is based on task demands, user capabilities, and system requirements. Meaning that during high task load conditions or emergencies the use of automation is increased and decreased during normal operations [3]. Another advantage of adaptive over static automation is the ability to reconstruct the task environment in terms of what is automated, how it is automated, what tasks may be shared, and when changes occur [2]. However, for an adaptive system to be efficient we require both the operator and the automated system to have sufficient knowledge of each other's current capabilities, performance and state [2]. As discussed by Byrne et al. (1996), there are three main approaches to address this issue, which either use model-based prediction or continuous measurements to determine the operators current state. In the scope of this thesis we will focus on the latter of the two. Usually, the way to measure human performance at work is to use physiological measures that reflect, more or less directly, aspects of brain functions. There is however a group of measures that is particularly favored among neuroergonomic researches. Those are the ones that are derived from the brain itself such as electroencephalography (EEG), magnetencephalography (MEG), and event-related potentials (ERPs), as well as measures related to the brain's metabolic and vascular responses such as positron emission tomography (PET), and functional magnetic resonance imaging (fMRI) [1]. Although, there are many advantages to this group of measures, such as the temporal resolution of ERPs and the spatial resolution of fMRI, there is still one major disadvantage. In

fact, the majority of the abovementioned measures are either too expensive, impose too much restrictions on the movement of the subject, or are simply unfit to be used in a portable system, which ultimately prevents their application in real-world settings. Also, the lack of comfort and therefore low operator acceptance of a certain measure could have detrimental effects on the success of an adaptive automated system.
An alternative could by provided by systems that use psychophysiological indices to trigger changes in automation. In general, pychophysiology is focused on physiological measures and their psychological correlates [1], but there are many psychophysiological indices that reflect underlying cognitive activity, arousal levels, and external task demand. Some of these include cardiovascular measures (e.g. heart rate, heart rate variability), respiration, galvanic response, ocular motor activity, and speech [5]. Even though, research examining the utility of psychophysiology in adaptive automation has been rare, physiological measures are likely to be considered in the design of adaptive systems, either in isolation or in combination with other measures [2]. In addition psychophysiological measures are usually well accepted, due to their non-invasive nature and easy of application.
In conclusion, human-machine interaction in highly automated workplaces can be optimized by creating a work environment that is sensitive to the mental state of human operators. Using psychophysiological measures, a constant feedback in the form of a neuroergonomic assessment of the operator condition is provided to the machine agent. Consequently, adaptation mechanisms can be deployed to alter the quantity, or quality of the workload according to operator capability.
Within the scope of this thesis we designed a wearable system to facilitate this process. We built our system upon the Empatica E4 wristband, a medical-grade wearable device capable of real-time physiological data acquisition. We then conducted a pilot experiment, deploying our system in near real-world conditions. We monitored 14 subjects performing two cognitive tasks, simulating high and medium workload conditions, and two sets of visual stimulation, designed to elicit specific emotional states. Finally, we evaluated different machine learning algorithms using the acquired dataset. With our system we address a distinct need for a reliable, and truly non-obstructive method of handling neuroergonomic assessment of human-machine interaction in collaborative work environments. Eventually allowing us to create integral workplaces, that are sensitive to a persons mental capability and capable of eliminating stress as one of the leading causes of injury and disease in the working population.

1.2 Theoretical Background

1.2.1 Psychophysiology

Psychophysiology is a field of study that investigates the relationship between reasoning, feeling, behavior, and the physiological correlates associated with them. Advances in neuroscience, endocrinology, immunology, and molecular biology led to great insights in the interdependence of physiological and psychological processes. The translation of psychological functional states, emotions, and behavioral patterns into physiological reactions and processes is essentially controlled by three different systems: the autonomic nervous system (ANS), the neuroendocrine system (NES), and the neuroimmunological system (NIS). A common use of psychophysiological measures is the study of physiological correlates of emotions, attention, stress, and other cognitive processes.

1.2.2 Psychophysiology in Adaptive Automation

To fully understand the role of psychophysiology in adaptive automation we will take a look at the theoretical framework behind it. However, providing a complete overview on this topic would be far too extensive for the scope of this thesis. Therefore, we will only give a short summary of the work done by Byrne and Parasuraman (1996).
The application of physiological measures in adaptive automation is built on the premise that there is indeed an ideal mental state for human operators in a given task environment and that any deviation from this state would be detectable in the measurement. This hypothesis is based on resource and capacity theories of information processing, which suggest that humans draw from a limited pool of resources whenever they process information [2]. Over the years, many researcher delivered evidence for a connection between this resource utilization and physiological measures of activation, therefore establishing the importance of psychophysiological measures in the field of adaptive automation.
However, psychophysiological measures perform a dual role in adaptive automation systems. First, there is the investigatory role, which is often referred to as the developmental approach. This approach is focused on using the information psychophysiological measures provide on the mechanisms underlying performance changes corresponding to changes in automation, and further the development of model-based and hybrid approaches [2]. The second role, is often characterized as the regulatory approach. Here, unique information about the human operator is gathered from psychophysiologic measurements. This information is then used as input to a hybrid adaptive logic, thus allowing for dynamic restructuring of the task environment. Although, this approach seems ideal to support the operation of an adaptive system due to its immediate effect on the automated work environment, there may be years of effort and considerable maturation in technology required for it to be efficient in its application.

1.3 Psychophysiological Measures

The identification of suitable psychophysiological measures plays a vital role to the success of an adaptive work environment. Considering the dual role framework of psychophysiology, there is a distinction to be made between the two applications in adaptive automation. Because the developmental approach is in alignment with the majority of applications in psychophysiological research, the often stated criteria of specificity, diagnosticity, and intrusiveness for selecting workload assessment techniques also hold for adaptive automation [2]. On the other hand, criteria for the regulatory role of psychophysiology in adaptive automation have to be more strict. As they become part of closed-loop systems operating in real-time their potential impact is far greater, and their effects more immediate compared to when used for developmental measures. In addition, the cost in terms of intrusiveness and technical requirements have to be weighed against the explanatory power of a certain measure. If the gain in predictive value does not offset the cost of implementation, a measure is not considered for applications outside of laboratory environment.
As the recent work is determined to employ the Empatica E4 wristband, we are limited to the measures that are provided by this platform. These measures are blood volume pulse (BVP), galvanic skin response (GSR), and skin temperature.

1.3.1 Photoplethysmography

Photoplethysmography PPG is an optical measurement technique, used to detect blood volume changes in the microvascular bed of tissue [6]. To work PPG only requires a few opto-electronic components. First, a light source is used to illuminate the tissue. Then a photodetector measures the variations in light intensity associated with changes in perfusion in the catchment area. The most common light sources in PPG produce wavelengths in the red or near infrared area. This specific part of the spectrum, also referred to as the optical water window, is chosen for its ability to pass through biological tissue with relative ease. Therefore, influences associated with light-tissue interactions are widely reduced and the measurement of blood flow or volume is facilitated at these wavelengths. Even so, because the PPG is representing an average of all blood volume in the arteries, capillaries, and any other tissue through which the light has passed. The PPG signal is dependent on the thickness and composition of the tissue beneath the sensor, as well as the position of the source in relation to the receiver of the infrared light [7].

The PPG waveform: characteristics and analysis

The PPG waveform is comprised of two major components. The pulsatile component, often referred to as the "AC" component, possesses a fundamental frequency of approximately 1 Hz, and it represents the increased light attenuation associated with the increase in microvascular blood volume with each heartbeat [6]. It is superimposed onto the much larger "DC" component, which relates to the tissue and the average blood volume contained in the observation area. Variations in the DC component are slower and caused by respiration, vasomotor activity and vasoconstrictor waves, as well as thermoregulation [6].

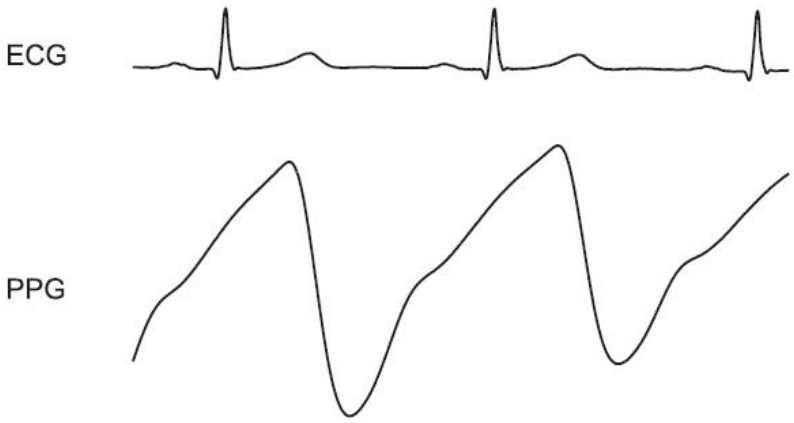


Figure 1.1: The PPG signal and the corresponding ECG. Displayed are the pulsatile AC component, which is super imposed on the much larger DC component. The PPG waveform represents light attenuation in relation to the blood volume in the tissue.

Its synchronization with the heart beat makes the AC pulse of the PPG waveform a valuable source of information on heart functions and condition. Based on the appearance of the AC pulse, two phases have been defined, reflecting its two most important properties. The first was labeled as the anacrotic phase and describes the rising edge of the pulse. This part of the waveform is primarily related to the systole. The second phase, shows the effects of diastole and wave reflections from the periphery of the vascular system. This phase is called catacrotic and can be observed in the successive falling edge of the pulse. In healthy subjects there usually is an observable dicrotic notch during this phase.

In addition to this coarse classification, a number of key landmarks have been defined to facilitate the analysis of the waveform and the underlying physiology. Depicted in 1.2 are the three main features that are derived from a single pulse. The pulse transit time to the foot (PTTf), and the pulse transit time to the peak (PTTp) are defined as the

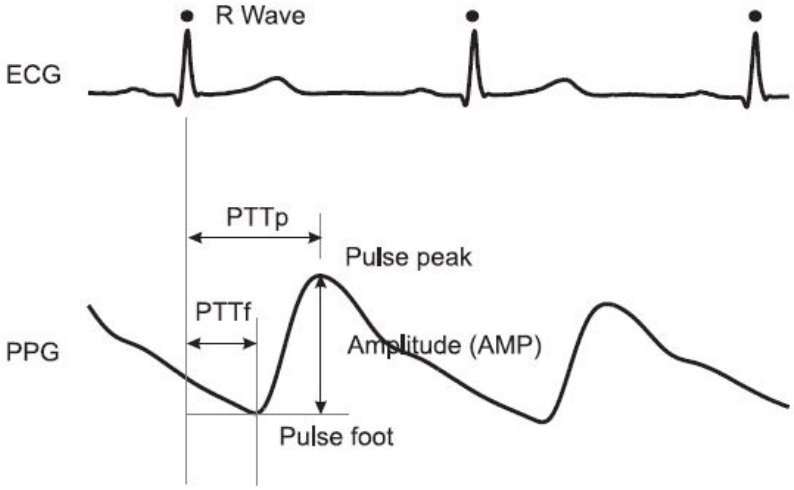


Figure 1.2: Characteristics of the PPG pulse waveform in relation to the ECG.

time delays between a heartbeat, indicated by the R-wave of the ECG, to the onset and the peak of the subsequent AC pulse. The amplitude of a pulse is determined by the absolute value of the displacement between its base and its peak, which are marked by the aforementioned temporal features.
However, the scale of these characteristics, as well as the overall appearance of the waveform are still subject to change. It is believed that these changes are largely caused by reflection of the pulse wave and the tapering down of the arteries towards the periphery [6].
Another important consideration in the analysis of PPG signals is their susceptibility to movement related artifacts. Although, there are a number of different artifacts that could occur we will only inspect the ones relative to our application. As we employed the Empatica E4, a wrist worn device, to measure BVP most of the artifacts are related to large movement of the arm and the wrist but also to small tremors in the hand or fingers. Additionally, extremes in physiological variation such as coughing and marked changes in the breathing pattern prove to be quite influential. Figure 1.3 shows an illustration by Allan (2007) depicting the effects of movement related artifacts to one minute PPG recordings that were taken at the index finger.

This concludes the section on the general waveform morphology and the reasoning for its variations. Lastly, we will take a closer look at the features we derived from the PPG measurement, specifically heart rate, and heart rate variability.

1.3.2 Heart Rate Variability

Variations in the length of the intervals between consecutive heart beats are called heart rate variability (HRV). Typically these inter beat intervals (IBIs) are determined by calculating the distance between two subsequent R-peaks of an ECG signal. However, they can also be derived from PPG signals. Again, IBIs are the time periods between the maxima of subsequent AC pulses. Since its first appreciation in 1965 HRV experienced a significant increase in popularity due to the apparent ease of derivation from widespread measures, such as ECG and PPG. In 1981, Akselrod et al. introduced power spectral analysis of HRV, contributing to the understanding of its autonomic background by relating the power content of certain frequency bands to sympathetic and parasympathetic activity [8]. In more recent years the increased interest in the application of psychophysiological measures in the field of adaptive automation led to an extensive discussion on HRV as a candidate measure. Byrne et al. (1996) argued in support of HRV as a possible index for both cognitive effort and compensatory effort. But, they also warned against the negligence of situation-related influences (e.g. the given task environment) in the process of signal interpretation. Thereby, a careless approach could have considerable consequences on the efficacy of adaptive automation.
In addition, the significance and meaning of the many different HRV measures are more complex than generally appreciated and consequently there is a high potential for incorrect conclusions and for excessive or unfounded extrapolation [8]. In 1996, the European Society of Cardiology and the North American Society of Pacing and Electrophysiology put together a Task Force to address this problem by developing appropriate standards for the acquisition and analysis of HRV. These standards remain valid until today and help preserving the integrity and reproducibility of HRV analysis if applied correctly.
Measures of the HRV can be divided into three general groups, the time domain methods, the frequency domain methods and non-linear methods.

HRV Measures

Of the three, time measures are perhaps the simplest to perform. These methods either determine the heart rate at any point in time or the intervals between successive normal complexes (in our case inter beat intervals) to calculate a series of statistical, and geometrical variables [8]. Statistical methods can be divided into two classes: those that are derived form direct measurements of inter beat intervals or instantaneous heart rate, and those derived from difference in inter beat intervals. Geometrical methods on the other hand, convert a series of IBI into a geometric pattern, such as the sample density distribution of IBI duration, sample density distribution of differences between adjacent IBI, Lorenz plot of IBI, etc., and then use a simple formula which judges the variability based on the geometric and/or graphic properties of the resulting pattern[8]. Figure 1.5 shows a variety of time-domain measures of HRV that have been recommended

by the Task Force of The European Society of Cardiology. Other common methods of HRV measures involve frequency-domain methods. By the means of power spectral density (PSD) analysis, these methods provide basic information on power distribution (i.e. variance) as a function of frequency. There are two ways to calculate PSD, non-parametric and parametric methods, both of which provide comparable results. In most cases non-parametric methods offer easier and faster calculation, whereas parametric methods, if applied correctly, can provide smoother spectral components which can be distinguished independently of pre-selected frequency bands. There are three main spectral components that are distinguished in a spectrum calculated from short-term recordings of 2 to 5 min: very low frequency (VLF), low frequency (LF), and high frequency (HF) components [8]. The distribution of the power and the central frequency of LF and HF are subject to changes in autonomic modulations of the heart period and therefore widely used measures in emotion recognition. Finally, we will have a look at non-linear methods. Although, the utility of these methods is still discussed by researchers, they are believed to provide valuable information for the physiological interpretation of HRV. Non-linear measures reflect the complex interactions of haemodynamic, electrophysiological and humoral variables, as well as autonomic and central nervous regulations involved in the genesis of HRV. At the moment, non-linear methods are speculated to be potentially promising tools for HRV assessment, but standards are lacking and the full scope of these methods cannot be assessed.

1.4 Machine Learning

Machine learning is a scientific study revolving around the development of algorithms and statistical models that provide computer systems with the ability to automatically learn and improve without being explicitly programmed. Machine learning techniques are commonly categorized as either supervised or unsupervised. This distinction is based on the type of input and output data they use as well as the type of problem they are intended to solve. Supervised learning algorithms require a set of data that contains both the inputs and the desired outputs to a certain problem. Based on this data, often referred to as training data, supervised algorithms derive a function that can be used to predict the output associated with new inputs. Supervised learning algorithms are generally used to solve classification and regression tasks. In contrast unsupervised learning algorithms are able to function on data sets that do not provide any output at all. They attempt to find a structure in the training data, by identifying commonalities. Based on the absence or presence of these similarities in new data they are then able to make predictions. Over the years a variety of algorithms have been developed for either category, each with its own objectives, strengths and weaknesses. Identifying an algorithm that is suitable for the desired task is one of the key components to a successful application of machine learning. Therefore, we will now take a closer look at the algorithm selection process used in the recent work.

1.4.1 Algorithm Selection

In the first step of the selection process we considered the work of previous researchers. In particular Wu et al. (2008), who presented the top 10 algorithms identified by the IEEE International Conference on Data Mining (ICDM) in December 2006 [9]. The algorithms were chosen according to the following steps. First they invited renowned researches of the field to each nominate up to 10 best-known algorithms in data mining. Each nomination had to provide the following information: the algorithm name, a brief justification, and a representative publication reference. After the nominations had been verified, those with less than 50 citations were removed. The remaining nomination were then organized in 10 topics: association analysis, classification, clustering, statistical learning, bagging and boosting, sequential patterns, integrated mining, rough sets, link mining, and graph mining. The final 18 candidates were then put to another vote with a much larger involvement of the research community. The results of this vote were the presented as the top 10 algorithms.
From this initial pool of algorithms we then selected those that were assigned to the topic of classification. As Python was the programming language we had agreed upon for all of our machine learning applications we selected the three algorithms that showed the highest compliance with the standard Python libraries from the subset of classification algorithms: kNN-classifiers, support vector machines, and ensemble methods. Also, we decided to include neural networks, in particular multi-layer perceptrons, into our final group of algorithms for exploratory reasons. In the following subsections we will provide a brief description of each of the final four machine learning algorithms.

1.4.2 k-NN Classifier

Nearest neighbors methods are possibly the simplest machine learning algorithms for supervised and unsupervised learning. The principle behind nearest neighbor methods is to find a predefined number of training samples closest in distance to the new point, and predict the label from these. The number of samples, named k, can be a user-defined constant (k-nearest neighbor learning), or vary based on the local density of points (radius-based neighbor learning). The distance can, in general, be any metric measure, but standard Euclidean distance is the most common choice. Neighbors-based methods are known as non-generalizing machine learning methods, since they do not attempt to construct a general internal model, but simply store instances of the training data. Classification can then be computed from a simple majority vote of the nearest neighbors of each point: a query point is assigned the data class which has the most representatives within the nearest neighbors of the point [10].

1.4.3 Support Vector Machine

Support vector machines (SVM) are considered one of the most robust and accurate methods among all machine learning algorithms. SVM have a sound theoretical foundation, require only a dozen examples for training, and are insensitive to the number of dimensions [9]. In a learning task with two classes, SVM aim to find the best classification function to distinguish between members of the two classes in the training data. The metric that is used to identify the best classification function can be realized geometrically. For a linearly separable data set, a linear classification function corresponds to a separating hyperplane $f(x)$ that passes through the middle of the two classes, separating the two. Once this function is determined, new data points x_n can be classified by simply testing the sign of the function $f(x_n)$; x_n belongs to the positive class if $f(x_n)$ is greater than zero. Additionally, SVM select the best classification function from all available hyperplanes by maximizing the margin between the two classes. The margin is defined as the amount of space, or separation between the two classes as defined by the hyperplane. Geometrically, the margin corresponds to the shortest distance between the closest data points to a point on the hyperplane [9]. The reason why maximum margin hyperplanes are emphasized this much in SVMs is that they provide the best generalization ability of the myriads of hyperplanes available. Therefore, they achieve the best classification performance on the training data, while still leaving enough room for the correct classification of the future data.

1.4.4 Decision Trees

Decision Trees are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning a hierarchic set of simple decision rules inferred from the data features. Training a decision tree for a certain problem is equal to learning a sequence of yes or no questions that will lead as fast as possible to the right answer. In machine learning these questions are known as tests. However, real life data is rarely available in the form of binary properties, but rather as continuous attributes that require tests in the form of inequations (e.g. is the attribute bigger or smaller than a certain value). To build the tree, the algorithm will go over every possible test and select the one that provides the most information on the target value. The first node, or the root, is representative of the entire data set. At each node a test will be administered and consequently the data set will be divided into two subsets of data, which again are represented by nodes. This process is then repeated, growing a tree of binary decisions until the data set is fully divided and the last nodes of each branch only contain values of one category. These final nodes are also called leaves. A leaf that only contains a single value is called pure. Predictions on a new data point can be made by examining the region of the attribute space that is associated with it and selecting the prevalent target value in this region.

This region can be located by simply reconstructing the tree. Starting from the root and following every decision until eventually one of the leaves is reached. Decision trees are easy to understand and to interpret. In addition they require very little data preparation, as opposed to other techniques, such as support vector machines. On the other hand, decision tree learners have a tendency to overfit. This means that they create overly complex trees that do not generalize well and therefore requires additional adjustments. There are two main strategies to prohibit overfitting in decision trees: Pre-pruning, which stops tree growth early on (e.g. limiting the maximum depth of the tree), and Post-pruning, which either removes or merges nodes that provide the least information.

1.4.5 Ensemble Techniques

Ensembles are methods that combine multiple machine learning models to create a new and more powerful model. Although there are many different ensemble-models, only two have proven effective on a wide variety of data sets: random forests, and gradient boosting machines. Both of these methods use decision trees as weak classifiers. Random forests are basically a combination of multiple, uniquely built decision trees. The basic principle of random forests is that every decision tree is inherently capable of making adequate predictions, but also overfits on some parts of the data set. And because all trees are unique, their overfitting is too. If enough of these trees are combined overfitting can be reduced, by taking the average of all the predictions. In contrast to this, the gradient boosting machines use trees, that are built successively. Each tree is designed to address the failures of its predecessor. Again, the basic idea is to use a number of partially well working decision trees and combine them. Gradient boosting often applies heavy pre-pruning to create rather flat trees with a maximum depth of one to five levels.

1.4.6 Multi Layer Perceptron

Multi-layer perceptron (MLPs) are a type of feed-forward neural networks, a family of learning algorithms inspired by the human brain. Similar to their biological role model, MLPs consist of a network of artificial neurons, called perceptrons, that are arranged in multiple layers. There are at least three layers in each MLP: an input layer, a hidden layer, and an output layer. Basically, MLPs are supervised learning algorithms that learn a function $f(\cdot) : R^m \longrightarrow R^o$ by training on a data set, where m is the number of dimensions for input and o is the number of dimensions for output. Given a set of features $X = x_1, x_2, ..., x_m$ and a target y, it can learn a non-linear function approximator for either classification or regression. The input layer is the leftmost layer of the network. It consists of a set of perceptrons $x_i|x_1, x_2, ..., x_m$ that each represent a single input feature. Next, in the middle section of the network, all of the hidden layers are located. Each perceptron of a hidden layer transforms all of the values from the previous layer with a

weighted linear summation $w_1x_1 + w_2x_2 + ... + w_mx_m$, followed by a non-linear activation function (e.g. the sigmoid function). The final layer is the output layer, which receives the values from the last hidden layer and transforms them into output value [10]. In MLP learning is achieved by adjusting the connection weights after each processing step. This adjustment is based on the size of error gained from comparing the output values with the expected results. MLP allow for complex models that capable of handling large data sets. However, a high complexity comes at the price of a high sensitivity to parameter changes, and data scaling methods, as well as longer training times.

1.5 Classification of Emotion

The means by which emotions can be distinguished from one another are called emotion classification. Over the last couple of decades this has been a strongly contested topic in emotion research and in affective science. One prominent approach to emotion classification, that has been widely accepted to this date, are dimensional models. Dimensional models of emotion attempt to characterize human emotions by determining their position in a two, or sometimes three dimensional space. After years of debate about the identity of these dimension two measures have claimed their place in most emotion models. Usually, the dimensions include some measure of valence or pleasantness and some measure of intensity or arousal [11]. For our work we employed the circumplex model by J. Russel (1980), one of the most prominent two-dimensional models for emotion classification. According to this model emotions are distributed in a two-dimensional circular space. The model space is represented by arousal (on the vertical axis) and valence (on the horizontal axis) and is centered on medium arousal and neutral valence. This allows for the possibility of emotions, or emotional stimuli that have high arousal and neutral valence (e.g. astonished, exited) [11] which separates the circumplex model from the rest of the two-dimensional models.

1.5.1 Emotion Elicitation

The American Psychological Association defines emotions as complex reaction patterns, involving experiential, behavioral, and physiological elements, by which an individual attempts to deal with a personally significant matter or event. The specific quality of an emotion, such as fear or shame, is determined by the specific significance of the event. For example, if the significance involves threat, fear is likely to be generated; if the significance involves disapproval from another person, shame is likely to be generated. Emotion typically involves feeling but differs from feeling in having an overt or implicit engagement with the world. This suggests that emotions can be elicited by using appropriate stimuli. In general, there are two types of stimuli in emotion elicitation are visual stimuli (e.g.

pictures or films) and acoustic stimuli (e.g. sounds, speech, music) that are widely accepted in psychology and emotion research. During our experiments we attempted to elicit emotions by displaying images, which is one of the most practiced methods in the elicitation of emotional and affective states. As stimuli we used a selected subset of the International Affective Picture System (IAPS).

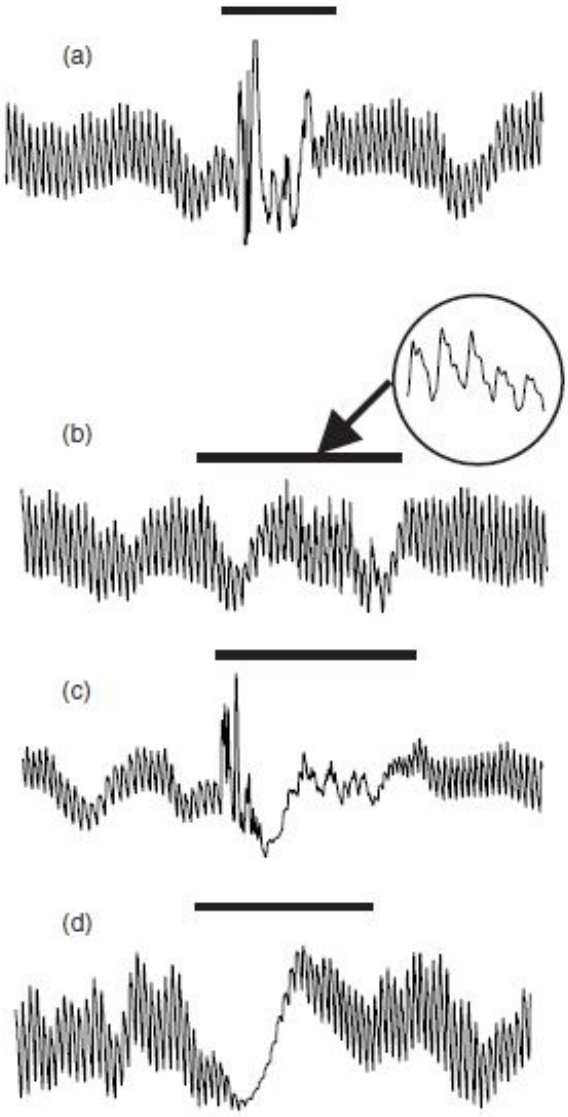


Figure 1.3: Examples of different types of measurement artifacts. All events were marked by black bars. (a) An episode of gross movement artifact of PPG probe cable tugging. (b) Hand or finger tremor. (c) a bout of coughing, and (d) marked changes in the breathing pattern (a deep gasp or yawn)

Variable	Units	Description
		Statistical measures
SDNN	ms	Standard deviation of all NN intervals.
SDANN	ms	Standard deviation of the averages of NN intervals in all 5 min segments of the entire recording.
RMSSD	ms	The square root of the mean of the sum of the squares of differences between adjacent NN intervals.
SDNN index	ms	Mean of the standard deviations of all NN intervals for all 5 min segments of the entire recording.
SDSD	ms	Standard deviation of differences between adjacent NN intervals.
NN50 count		Number of pairs of adjacent NN intervals differing by more than 50 ms in the entire recording. Three variants are possible counting all such NN intervals pairs or only pairs in which the first or the second interval is longer.
pNN50	%	NN50 count divided by the total number of all NN intervals.
		Geometric measures
HRV triangular index		Total number of all NN intervals divided by the height of the histogram of all NN intervals measured on a discrete scale with bins of 7·8125 ms (1/128 s). (Details in Fig. 2)
TINN	ms	Baseline width of the minimum square difference triangular interpolation of the highest peak of the histogram of all NN intervals (Details in Fig. 2.)
Differential index	ms	Difference between the widths of the histogram of differences between adjacent NN intervals measured at selected heights (e.g. at the levels of 1000 and 10 000 samples)[123].
Logarithmic index		Coefficient φ of the negative exponential curve $k \cdot e^{-\varphi x}$ which is the best approximation of the histogram of absolute differences between adjacent NN intervals[123].

Figure 1.4: Summary of recommended time-domain measures [8].

Variable	Units	Description	Frequency range
		Analysis of short-term recordings (5 min)	
5 min total power	ms^2	The variance of NN intervals over the temporal segment	approximately $\leq$0·4 Hz
VLF	ms^2	Power in very low frequency range	$\leq$0·04 Hz
LF	ms^2	Power in low frequency range	0·04–0·15 Hz
LF norm	n.u.	LF power in normalised units LF/(Total Power–VLF) × 100	
HF	ms^2	Power in high frequency range	0·15–0·4 Hz
HF norm	n.u.	HF power in normalised units HF/(Total Power–VLF) × 100	
LF/HF		Ratio LF [ms^2]/HF [ms^2]	
		Analysis of entire 24 h	
Total power	ms^2	Variance of all NN intervals	approximately $\leq$0·4 Hz
ULF	ms^2	Power in the ultra low frequency range	$\leq$0·003 Hz
VLF	ms^2	Power in the very low frequency range	0·003–0·04 Hz
LF	ms^2	Power in the low frequency range	0·04–0·15 Hz
HF	ms^2	Power in the high frequency range	0·15–0·4 Hz
α		Slope of the linear interpolation of the spectrum in a log-log scale	approximately $\leq$0·04 Hz

Figure 1.5: Summary of recommended frequency-domain measures [8].

2 Problem Analysis and Goals

2.1 Related Work

In the past a substantial part of research in the field of Human Machine Interaction has been focused on the use of emotion recognition to give robots the ability to perceive and appropriately react to human emotions. With the appreciation of psychophysiological signals as potent markers of emotional states, a new wave of studies has been conducted on their deployment in the recognition process of adaptive human machine interaction. We will now have a look at three studies that are closely related to ours. In this process we will investigate key aspects and show potential problems. The primary goal of all studies was to distinguish emotional states in subjects using a combination of psychophysiological measures and some machine learning technique. Measures of heart rate such as BVP or ECG, as well as galvanic skin response were used in all three as physiological indicators of emotion. In addition, respiration and EMG were used by both Maaoui and Pruski (2010) and Picard et al. (2001). Lastly, Maaoui and Pruski (2010) as well as Kim et al. (2004) employed skin temperature in their work. All measures were applied in their default measurement locations. Although, this might be acceptable for scientific settings in a lab, we argue that some of these location need to be adjusted or entirely ruled out to provide the minimum comfort to be applicable in a real-world setting. Especially, the negative effects of facial EMG, as well as the finger sensors used to measure GSR and BVP to user acceptance are easily comprehensible in an work environment where workers are physically engaged. Also, most of the sensors that were used in these studies still relied on a wired connection to deliver their data to a main processing device. A fact that further reduces practicability and in some workplaces may even increase the risk of injury. Proceeding to the classification of emotion, we find that all three studies reported above average results. A classification accuracy of 81% was achieved by Picard et al. (2001) while using Sequential Floating Forward Search and Fisher Projection methods to classify a total of 8 different emotional states. They used data they collected from a single subject over a time period of 30 days, with an average recording time of 25 minutes per day. Although, this method may account extremely well for daily changes in the recorded signal, it creates a classification that is strongly dependent on a single subject. The customization of an emotion recognition system to a single user may be of interest in the final stages of implementation, but one could argue that at this stage of research a user-independent system is far more beneficial due to its universal applicability. Next, we will consider the work of Maaoui and Pruski (2010) who achieved remarkable results.

They were able to reach a classification accuracy of 92% over 6 emotions by using SVM and Fisher Linear Discriminant analysis. However they used a set of elaborate protocols for emotion induction in combination with the IAPS and an extensive lab setting, which would again not be feasible to use in an adaptive automation workplace. Lastly, Kim et al. (2004) who attempted a user independent emotion detection system, using visually and acoustically stimuli in combination with the IAPS to elicit specific emotion in large groups (group 1: n = 125, group 2: n = 50) of children below the age of 9. Although, this strategy might be able to elicit stronger emotions that are much better represented by physiological correlates, it could be argued to be counter productive towards the generalization ability of their machine learning algorithm in an adult target population.

2.2 Problem Oriented

Section 2.1 shows that regardless of the significant accomplishments and overall advance of the field, there are still problems that are left to be resolved. As we pointed out earlier, although there were some attempts to conduct emotion recognition experiments in more realistic scenarios, the recording systems, and most of all the recording techniques (i.e. sensor locations) remained widely unimproved. Therefore, we will stray away from this pattern and use the Empatica E4, a wrist-worn system that is capable of wireless data transmission and features a single measurement site for all sensors. We believe that this is the right step towards truly unobtrusive emotion recognition allowing the field to advance by enabling more realistic experimental settings. Another important aspect we gathered from our literature research is the apparent focus on detecting only pure emotional states. Although, this may be useful in some applications, emotions are rarely presented in this form under realistic circumstances. Considering the setting of an adaptive automation workplace we will attempt to elicit and classify more generalized mental states (e.g. is the subject feeling pleasant or unpleasant?). Further, we will create our own database for classification including multiple subjects, representative of a working population to attempt a user independent system. Regarding the selection of psychophysiological measures we will comply with the studies we have presented in 2.1. However, as we have only access to a subset of these measures, we will use what is available to us: BVP, GSR, and skin temperature.

2.3 Aim, Objective, and Scope

As mentioned before, the aim of this thesis was to facilitate the neuroergonomic assessment of human robot interaction based on the real-time measurement of psychophysiological signals using the Empatica E4 wristband. Meaning, we are faced with the daunting task

of developing a compact emotion classification system that is capable of consistently gathering high quality data and reliably performing classification based on meaningful features. Therefore, we focused heavily on the development of such a system, limiting our scope to the engineering, and assembling of all system components and the final evaluation, based on system performance.
In the following we listed all the milestones that were necessary to complete this task.

MS 1: Development of a data extraction method that provides for real-time access on the raw signal data, as the E4 is actually designed for downstream data analysis only.

MS 2: Construction of a signal processing pipeline that compensates for artifacts while upholding signal integrity for feature extraction.

MS 3: Building a database for machine learning using authentic data gathered with our system in a series of measurements.

MS 4: Evaluation of system capabilities using different machine learning techniques.

2.4 Research Questions

RQ 1: Is it possible to get real-time access to the data we record with the Empatica E4?

RQ 2: Is it possible to detect and distinguish different degrees of mental workload, as well as two emotional states, such as pleasant and unpleasant, using common machine learning techniques on three physiological signals that were recorded with the Empatica E4 in an experimental setting?

RQ 3: Which algorithm is best suited for the classification task introduced in RQ2?

2.5 Research Methodology

In this section, we present the methodology with which we attempted to solve our research questions.

RQ 1: Is it possible to get real-time access to the data we record with the Empatica E4?

We approached RQ1 by researching the E4's data streaming functionality. We discovered that there was a possibility to transmit raw-data from the E4 to a Windows computer

by using a streaming server application in combination with a specific USB Bluetooth receiver. The streaming server App is a developer tool provided by Empatica and can be used to register and pair, one or more Empatica devices to a certain PC. Once the pairing process is completed it is possible to access the data stream of a paired E4 device using a TCP client. We then conducted a web research on past projects involving data transmission to the Empatica streaming server. Among others we found the Empatica BLE Client for Matlab environment, developed at ICAT Virginia Tech. This TCP client was capable of storing the raw-data it received from the Empatica E4 in text files of the CSV-format. Based on these findings, we were confident to build our own TCP client, capable of providing real-time access to the raw signal data, using the Empatica web resources for developers (e.g. Documentation on message protocol, data streaming packets, and the E4 streaming server) in combination with the MATLAB environment.

RQ 2: Is it possible to detect and distinguish different degrees of mental workload, as well as two emotional states, such as pleasant and unpleasant, using common machine learning techniques on three physiological signals (BVP, GSR, and skin temperature) that were recorded with the Empatica E4 in an experimental setting?

Answering RQ2 required us to gather relevant data. We therefore conducted experiments in both, the Mindscan Lab at HTW Saar, as well as the Green Lab at the University Hospital Saarland. The recorded data was then used to evaluate the different machine learning algorithms we selected following the procedure we already described in 1.4.1.

RQ 3: Which algorithm is best suited for the classification task introduced in RQ2?

The answer to RQ3 is based on the outcome of RQ2 and an evaluation of its results.

3 Materials and Methods

In this chapter, we present all the hardware components and software applications that were used in the process of this thesis.

3.1 Hardware

3.1.1 Empatica E4

The Empatica E4 wristband is a wearable wireless device designed for comfortable, continuous, real-time data acquisition. It is a class IIa medical device in the EU, according to CE Cert. No. 1876/MDD (93/42/EEC Directive) and was designed for daily life usage [12].

Figure 3.1 shows an overview of the entire E4 wristband from either side indicating key attributes as wells as a total of four different sensors that will be discussed briefly in the following:

- **Photoplethysmography (PPG)** to provide blood volume pulse (BVP), from which heart rate, heart rate variability and other cardiovascular features may be derived
- **Electrodermal activity (GSR)** is used to measure sympathetic nervous system arousal and to derive features related to stress, engagement and excitement
- **3-Axis accelerometer** to capture motion-based activity
- **Infrared thermopile** for reading skin temperature

As the E4 is intended to be worn on the wrist these sensors are set up in a specific way to provide for optimal use. As can be seen on 3.1 the majority of the sensors are located on the backside of the main unit not including the GSR-sensor, which is located on the wristband itself.
Wearing the E4 wristband is equally intrusive to wearing a watch and therefore providing

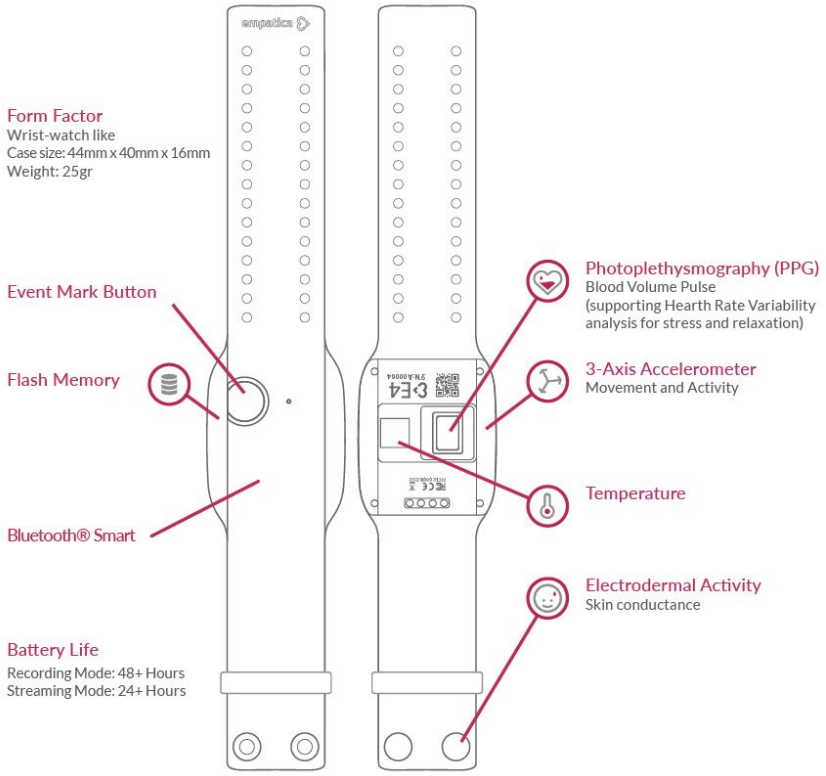


Figure 3.1: Overview of the Empatica E4 wristband.

a high level of convenience compared to other physiologic measures such as electrocardiogram ECG or electroencephalogram EEG.

Sampling Specifications

All recordings were performed using only software licensed by Empatica. Using the approved streaming server application and the compatible Bluetooth dongle, the recorded data was streamed directly to an operator's personal computer via a Bluetooth connection.

EDA sensor

- Sampling frequency: 4 Hz (Non customizable).
- Resolution: 1 digit 900 pSiemens.

- Range: 0.01 μSiemens – 100 μSiemens.
- Alternating current (8Hz frequency) with a max peak to peak value of 100 μAmps (at 100 μSiemens).
- Electrode(Placement): on the ventral (inner) wrist.
- Electrode(Build): Snap-on, silver (Ag) plated with metallic core.
- Electrode(Longevity): 4–6 months

PPG sensor

- Sampling frequency 64 Hz (Non customizable).
- LEDs: Green (2 LEDs), Red (2 LEDs) Photodiodes: 2 units, total 15.5 mm^2 sensitive area.
- Sensor output: Blood Volume Pulse (BVP) (variation of volume of arterial blood under the skin resulting from the heart cycle).
- Sensor output resolution 0.9 nW / Digit.
- Motion artifact removal algorithm: Combines different light wavelengths. Tolerates external lighting conditions.

Infrared Thermopile

- Sampling frequency: 4 Hz (Non customizable).
- Range(Ambient temperature): -40...85degC (if available).
- Range(Skin temperature): -40...115degC.
- Resolution: 0.02degC.
- Accuracy $\pm$0.2degC within 36-39degC.

Real-time clock

- Resolution(Recording mode): 5s synchronization resolution. Average of 6 seconds in 6 million seconds drift.
- Resolution(Streaming mode): Temporal resolution up to 0.2 seconds with connected device.

3.1.2 Processing Unit

The processing unit it the expression given to the computer, which is used to run any software that is essential for data transmission and data processing, such as the Empatica streaming server application, MATLAB, and PyCharm. We used a Lenovo ThinkPad, with an Intel(R) Core(TM) i5-6200U, CPU @ 2.3 GHz, 8 GB RAM, running a 64-Bit version of Windows 10.

3.2 Software

3.2.1 E4 Streaming Server

The E4 streaming server for Windows (version of May,2018) is a software application that allows to forward real-time data of one or multiple Empatica E4 devices to one or multiple TCP socket connections. However, as each TCP connection is limited to receiving data from only one Empatica E4 device, the connection to multiple devices would also require multiple TCP connections. The E4 streaming server is intended to provide access to the data streams using scripts and applications of any programming language [13].

3.2.2 MATLAB

MATLAB is a computing and visualization software package, published by MathWorks. It combines a desktop environment tuned for iterative analysis and design processes with a high level programming language for matrix-based mathematics. We used MATLAB (version R2015a) to create a TCP client that was deployed in our data extraction pipeline to send commands to, and receive messages from, the E4 streaming server. Thereby,

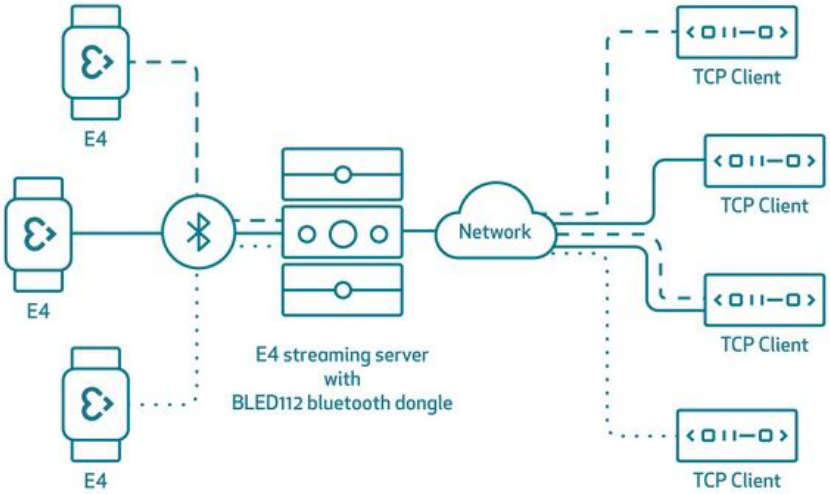


Figure 3.2: Illustration of the connectivity and function of the E4 streaming server. On one side are E4s, connected over BTLE to the E4 streaming server using the BLED112 dongle. On the other side are TCP clients, connected to the E4 streaming server through TCP connections over the network. The lines originating from the E4s illustrate the data flow from the E4 through the E4 streaming server to the subscribed TCP client. For example, the data from the first E4 is forwarded by the E4 streaming server to the first and third TCP client. [13]

allowing us to connect an E4 device and control data transmission via a simple user interface.

3.2.3 PyCharm

PyCharm is an integrated development environment (IDE) for the Python programming language. It is developed by the company JetBrains and provides easy access to a large collection of scientific tools, used for data analysis and visualization. We used PyCharm to manage all data related tasks, such as pre-processing, feature extraction, and machine learning implementation.

3.2.4 PsychoPy

PsychoPy is an open-source package for running experiments in Python. PsychoPy combines the graphical strengths of OpenGL with the easy Python syntax to give

scientists a free and simple stimulus presentation and control package. It is used for psychophysics, cognitive neuroscience and experimental psychology. We used PsychoPy to create and display the paradigm for our experiments.

3.2.5 IAPS

The International Affective Picture System (IAPS) is being developed to provide a set of normative emotional stimuli for experimental investigations of emotion and attention. The goal of the IAPS is to develop a large set of standardized, emotionally-evocative, internationally-accessible, color photographs that includes contents across a wide range of semantic categories. The IAPS is being developed and distributed by the Center for Emotion and Attention (CSEA) at the University of Florida [14].

4 Experimental Work

4.1 Pilot Experiment

In this section, we introduce the pilot experiment we conducted prior to the large scale experiment. In general pilot experiments are a method to test feasibility and provide experience-based guidelines for follow up experiments.

4.1.1 Introduction

The pilot experiment was conducted in the facilities of Systems Neuroscience and Neurotechnology Unit, particularly the Mindscan Lab, located at the HTW Saar (Technikum). The main goal of this experiment was to gather an initial set of authentic data while also testing the functionality of our data transmission routine and familiarizing ourselves with hardware and software behavior under experiment conditions.
The measurement was performed in a closed off laboratory environment with controlled temperature and illumination. Physiological signals were recorded using the Empatica E4, which was placed on the wrist of the non-dominant hand, and a laptop, serving as Processing Unit, that was placed approximately 1.5 meters away from the subject position. We measured 2 subjects in this pilot experiment. They were bachelor level male students, studying at the HTW Saar aged between 26 and 30 years.

4.1.2 Objectives

- Test the procedure and the experimental design
- Identify problems and optimization possibilities
- Estimation of time and resource requirements
- Inspect data quality

4.1.3 Setup

The experimental setup can be divided into two major components. The first component, is responsible for data acquisition and is comprised of the Empatica E4 and the Processing Unit. Whereas the second, handles the presentation of the experimental paradigm using a second Notebook in combination with a monitor. During the experiment the subject is positioned in a chair in front of the monitor. The recording site is surrounded by movable walls serving as visual cover. Directly behind the cover, and to the side of the recording site, a desk has been placed to provide space for the Processing Unit, the operator, as well as any additional devices needed for the experiment. The distance between the E4, when worn by the subject, and the Processing Unit was approximately 1.5 meters. The room was calm and dimly lit, providing enough light to allow the participants to comfortably observe the paradigm on the monitor without any reflections, or being blinded by incident sun light. The temperature was regulated by an air conditioning system and kept at a constant level during the experiment to negate environmental influences.

4.1.4 Procedure

In preparation of the experiment all participants were given a step-by-step explanation of the experimental procedure, to make sure they were in a relaxed and comfortable mood. After, the subjects had given their consent, they were asked to remove any electronic devices from their person and take a seat in the chair. Then, they were instructed to put on the Empatica E4. If necessary an operator would assist them. Once they were finished, the operator inspected the placement of the wristband in regards to the general fit, and sensor-skin contact. Prior to the start of the paradigm the participants were again reminded to be as relaxed as possible, make themselves comfortable, and avoid extensive movement. During the experiment the subjects were required to sit in front of the monitor and observe the paradigm, which was displayed on it, for a period of 5 minutes.

4.1.5 Paradigm

The paradigm that was used in the pilot experiment, was comprised of a single picture showing a simple black cross-hair on a white background. The paradigm was created with PsychoPy, which was running on a Notebook, connected to the monitor.

4.1.6 Problem Oriented

From the feedback we were given by the subjects, we were able to identify the following problems:

- The participants reported that the motive of the paradigm was causing them discomfort as the contrast between black and white was too extreme.
- Both participants felt that the chair was not providing enough comfort due to missing armrests and height difference to the table.

Further, we observed a slight offset in system time between the Processing Unit and the Notebook we used to run PsychoPy that could potentially raise issues in subsequent data analysis.

4.1.7 Solution Oriented

To address the aforementioned problems we altered the color scheme of our paradigm to provide a softer contrast (grey/white), exchanged the chair, and implemented a time synchronization routine in the experimental procedure.

4.1.8 Results and Conclusion

Because the results of our pilot experiment were overall positive we felt reassured to proceed to the main experiment with few alterations. We were able to successfully test the experimental procedure and the functionality of the setup. Further we were able to record quality data and gained valuable insight on the extent of the pre-processing needed to prepare the signals for feature extraction. We also gained a better understanding of the time-frame needed for the preparation, and the post-production of the experiment.

4.2 Main Experiment

In this section we present the main experiment of this thesis, which was conducted subsequently to our pilot experiment.

4.2.1 Introduction

In the second experiment, we measured a larger group of subjects, aiming to collect a reasonable amount of data for machine learning. Therefore, the fundamental idea of this experiment was to elicit several emotional states in the participants by using a combination of cognitive tests and visual stimulation. Similar to the pilot experiment, we conducted prior, all measurements took place within the facilities of Systems Neuroscience and Neurotechnology Unit, particularly the Green Lab, located at the University Hospital Saarland, and the Mindscan Lab, located at the HTW Saar (Technikum). The sessions that took place at the Mindscan Lab, used the exact setup from the pilot measurement except for a few modifications to account for the problems that were identified in 4.1.6. The Green Lab setup also featured a slightly modified illumination system to provide for similar circumstances in both labs.

4.2.2 Participants

In total 14 subjects, of which 7 were male and 7 female, participated in the experiment. Subject age ranged from 24 to 50 years, with an average of 29 years. At the day of the experiment all participants reported to be feeling well and capable of partaking in the experiment.

4.2.3 Methods

In the following we will briefly discuss the methods we used in our experiment to elicit certain emotional states, as well as different degrees of mental workload.

Mental Arithmetic Stress Test

The forced mental arithmetic test is one of the most widely used techniques for the elicitation of physiological arousal and cardiovascular activation [15]. Jern et al. (1991) introduced a version of this method that required subjects to perform forced mental arithmetic for 10 minutes with serial subtractions of 7 from 700, while trying to keep pace with a metronome at a rate of approximately 90 beats/min. Since none of the subjects was capable of matching this speed for more than a brief period of time, they were asked repeatedly to increase their speed by short vocal comments. Also, every time a subject would give a wrong number, they were immediately corrected. Throughout, the test subjects were encouraged to perform at their maximum speed, without harassment

and in an emotionally neutral attitude. We used a slightly modified version of this test to induce a high amount of stress in our subjects. Similar to Jern et al. (1991) we had our subjects perform serial subtractions of 7 starting from 700 for 5 minutes. Also, we substituted the metronome with a visual stimulus, in the form of a blinking dot, on a screen in front of the subject. The dot would flash at a speed of approximately 60Hz and change color depending on time passed: green (t < 180s), yellow (180s < t < 240s), and red (t >= 240). The operator monitored the process and intervened when one of the following situations occurred:

- the subject made a mistake → the operator would give the last correct value, ask the subject to continue
- the subject slowed down significantly → the operator would remind the subject to speed up
- the subject completed the task before the 5 minute mark → the operator would ask the subject to continue subtracting from 1400

The test was followed by a short questionnaire, asking subjects to give a subjective rating of the test difficulty and their current stress level, and a resting period.

Stroop Test

The Stroop Test has a lengthy history as an experimental measure in psychological studies and more recently has been adapted for clinical neuropsychological use [16]. It measures the relative speed of reading names of colors, naming colors, and naming colors used to write an incongruous color name (e.g. the color blue is used to write the word red). The last task requires subjects to override a reading response. This conflict interference situation is called the Stroop Effect [16]. We used a simplified version of this test to cause low to medium levels of stress in our subjects. The subjects were placed in front of a computer monitor and asked to determine the color of a color-word, which was displayed in the center of the screen using PsychoPy. They had to select the correct answer from two possible choices, which were displayed to either side of the color-word, by pressing certain keys on a keyboard. Each color-word constituted one trial. In total, the subjects had to complete 220 trials, featuring 11 different colors. When compared to other versions of the test, this amount of trials may seem extensive but it was derived from a minimum test length of 5 minutes, as well as an estimated average response time of 800ms per trial. Also, the color palette was extended to prevent repetition among the trials. Afterwards, subjects were asked to answer a short questionnaire identical to the one in 4.2.3.

Visual Stimulation

Visual stimulation is one of the most common methods used to elicit emotional states in psychological research. We used a collection of selected pictures from the IAPS to elicit two different emotional states in our subjects: (a) pleasant and (b) unpleasant. For (a) we selected a set of 32 pictures with positive, joyful, and few neutral motives, whereas for (b) we put together a set of 30 images with overall negative motives, aiming to induce feelings of unease, tension, and slight disgust. During the experiment subjects were required to complete one stimulation session, as well as one relaxation session, which each lasted approximately 5 minutes, for both emotional state. Each stimuli was presented for a duration 7 seconds followed by a neutral image with a cross-hair in the center of the screen, which was used to reset before the next stimulus. After each stimulation session, subjects had to rate the pictures using the circumplex model, mentioned in 1.5. Therefore, each picture of the set was presented anew, next to a scale for each dimension. The arousal scale ranged from 1 (very low) to 10 (very high), and the valence scale ranged from 1 to 7, with 1 to 3 representing different degrees of negative valence, 4 being equal to 0 valence (i.e. neutral), and 5 to 7 representing different degrees of positive valence.

4.2.4 Experimental Design

The general design of the experiment was as follows. After an initial baseline measurement of 5 minutes, two mental stress tests (Arithmetic stress test, and Stroop test), and two sessions of visual stimulation were performed. Each segment of the experiment was followed by a 5 minute relaxation interval (cooldown sessions). Both, the starting and the ending time of each session were saved automatically. In addition, an operator was present throughout the entire experiment, monitoring system performance, controlling the paradigm, and providing assistance if necessary.

4.2.5 Procedure

Upon reporting to the lab, the subject was asked to remove all electronic devices and then placed in the examination chair. The experiment begun with a short briefing session of approximately 5 to 10 minutes, in which the subjects were given a coarse outline of the experiment, as well as a short questionnaire on personal information, such as age and gender. After the subject confirmed that he/she was feeling capable of partaking in the experiment, the Empatica E4 was put on the wrist of the non-dominant hand of the subject. If necessary the operator would assist in the process to achieve a comfortable fit. To conclude the preparation phase the operator inspected sensor placement and tested system functionality with a 1 minute test measurement.

Once the functionality was confirmed, the experiment commenced following the order described in 4.2.4. Whereby, each session was initiated with a step-by-step explanation of the task, as well as an instruction to the subject's responsibilities (e.g. operating the keyboard, relaxing during cooldown sessions). The paradigm was mostly controlled by the operator, except for the ratings at the end of a session, and the Stroop test (both required the subject to use mouse and/or keyboard). After the last session was completed, the paradigm was terminated, the measurement was stopped, and the E4 was removed. The experiment was then concluded by a debriefing session of roughly 5 minutes, giving subjects the opportunity for questions and feedback.

5 Data Analysis and Interpretation

The signals we obtained from the experiment were analyzed and interpreted using the methods presented in this chapter. Starting with specifics of the data collection process, we continue by discussing signal processing algorithms and feature extraction, and end with details on classification procedure.

5.1 Data Collection

As mentioned in section 1.3, the psychophysiological measures, selected for this thesis, were blood volume pulse (BVP), galvanic skin response (GSR), and skin temperature. These signals were recorded using the integrated Photoplethysmography (PPG), Electrodermal Activity (EDA), and Infrared Thermopile sensors of the Empatica E4 wristband. All recordings were conducted using a sampling frequency of 64Hz for BVP, and 4Hz for both GSR, and skin temperature. From the 14 subjects that partook in our experiment we obtained 14 sets of physiological data. The data of one subject was removed due to substantial artifact contamination, resulting in a total of 13 sets of physiological data that were used in the subsequent process. Each set was comprised of roughly 80-90 minutes of recordings, which complied to approximately 326.400 samples for BVP, and 20.400 samples for GSR, and skin temperature. The collected data was continuously streamed to a TCP client, running on the nearby Processing Unit, via a Bluetooth connection and saved into csv files after every full minute of recording. The saved files contained data samples of all three data streams and were named following the pattern shown in 5.1.

$$\text{recording\_yyyy-mm-dd-hh-mm-ss} \tag{5.1}$$

Then again, each data sample was comprised of three components which identified the data stream, the sample time, and the sample value. An example of this pattern is shown in 5.2 for a sample of the BVP signal stream.

$$\text{E4\_Bvp } 1569592961{,}01857\ 25{,}33795 \tag{5.2}$$

5.2 Data Preparation

As the data was saved in a number of individual csv files the goal of the initial step of data processing was dedicated to translate the data samples into a usable format and separate the individual data stream (i.e. BVP, GSR, and skin temperature). This goal was achieved using two simple python scripts. The first program would locate all csv files in a certain folder and seamlessly merge them to a single csv file. Afterwards, the second script was used to scan all entries of this new file and sort them by their stream tag, which was indicated by the first portion of each data sample (see 5.2), using a simple string comparison method. This process resulted in a total of 6 data files (i.e. 2 files per data stream), of which each file contained either the sample values or the sample times of a single data stream. This format made the data easily distinguishable and accessible for the final step of data preparation, which was data segmentation. We divided the data streams into segments that were associated with the individual sessions of the experiment, using automatic timestamps that were generated during the experiment. The extracted segments were then saved in individual files that were appropriately named to provide explicit information about their content to facilitate subsequent processing steps. The naming pattern is shown in 5.3.

_DataStream_Datatype_SubjectID_SessionLabel_StartTime_EndTime (5.3)

5.3 Data Analysis

All aspects of data analysis were handled using the programming language Python in combination with the PyCharm IDE. Therefore, this section is dedicated to explaining the general strategy, as well as the scripts that were built in the process of managing data processing, and feature extraction tasks.

5.3.1 BVP

Pre-Processing

The main objective of BVP pre-processing was the detection of AC pulses, or α waves, in the BVP signal. To guarantee high quality peak detection even under challenging conditions, we implemented an algorithm that was introduced by Elgendi et al. (2013) for this very reason. The peak detection algorithm is based on event-related moving averages with dynamic thresholds, and is comprised of three main stages: pre-processing

(bandpass filtering and squaring), feature extraction (generating potential blocks using two moving averages), and classification (thresholding)[17].

The following paragraph will elaborate on the individual steps of the algorithm.
Bandpass Filtering. A zero-phase second-order Butterworth filter, with a bandpass of 0.5-8 Hz was implemented to remove baseline wander and non-peak-related high frequency components. The filter output was then applied to the raw PPG signal resulting in the filtered signal $S[n]$.
Clipping. In preparation of the next step of the algorithm, the filtered signal was clipped by removing the signal below zero. This resulted in the clipped signal $Z[n]$.
Squaring. Squaring was used to emphasize large differences resulting from the systolic waves, while simultaneously suppressing smaller differences caused by diastolic waves and noise [17]. This resulted in the signal $y[n]$ which is equal to $Z[n]^2$
Generating Blocks of Interest. Blocks of interest are generated using the two event-related moving averages MA_{peak} and MA_{beat}. MA_{peak} was used to mark systolic peak areas in $y[n]$ and is given by the equation 5.4. Where W_1 represented the window size of the systolic peak duration and was set to a value of 111 ms.

$$MA_{peak}[n] = \frac{1}{W_1} \cdot (y[n - \frac{W_1 - 1}{2}] + ... + y[n] + ... + y[n + \frac{W_1 - 1}{2}]) \tag{5.4}$$

MA_{beat} was used to mark heartbeat areas and given by the equation 5.5. Where W_2 represented the window size of one heartbeat duration and was set to a value of 667 ms.

$$MA_{beat}[n] = \frac{1}{W_2} \cdot (y[n - \frac{W_2 - 1}{2}] + ... + y[n] + ... + y[n + \frac{W_2 - 1}{2}]) \tag{5.5}$$

Thresholding. The first dynamic threshold THR_1 was used to mark potential peak areas in $y[n]$ by creating a block of interest for every interval where THR_1 was greater than MA_{peak}. THR_1 was derived from MA_{beat} following equation 5.6.

$$THR_1 = MA_{beat}[n] + \alpha \tag{5.6}$$

$$\alpha = \beta \cdot \bar{z} \tag{5.7}$$

Where β was set to a constant value of 0,02 and $\bar{z}$ was the statistical mean of $y[n]$. Figure 5.1 demonstrates the idea of using two moving averages to generate blocks of interest.

Although this process generated many blocks, only some of them would contain the desired feature (i.e. the systolic peak). Therefore, it was necessary to identify and reject the blocks that resulted from diastolic waves and noise. The rejection was based on the second threshold THR_2, which corresponded to the anticipated width of a systolic peak and was set to W_1. Whenever a block was wider than or equal to THR_2 it was classified as valid.
Peak Detection. In the last step of the algorithm systolic peaks were detected by localizing the maximum absolute value in valid blocks of interest.

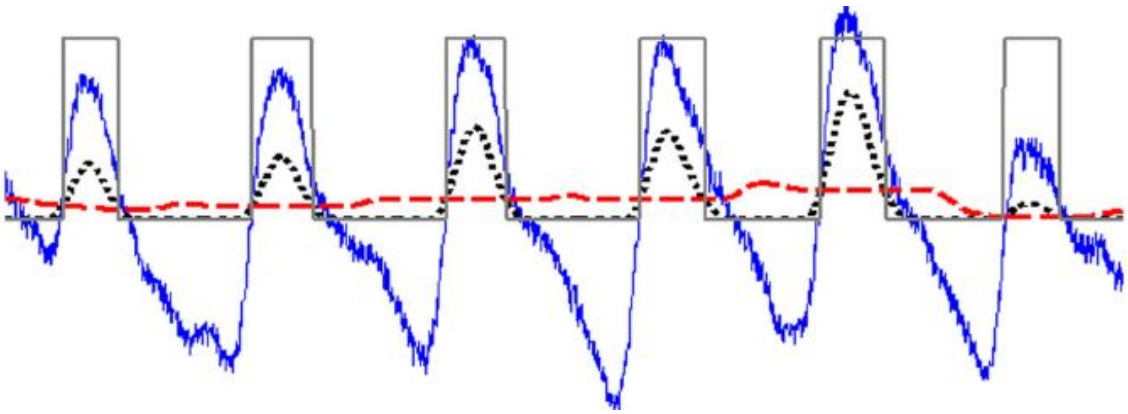


Figure 5.1: Blocks of interest are shown as grey squares over the filtered signal (in blue). The blocks are generated, using the two moving averages MA_{peak} and MA_{beat}, which are depicted as a black dotted signal and a dashed red signal respectively. (source: doi:10.1371/journal.pone.0076585.g009)

Inter-Beat-Intervals

Using the indices of valid systolic peaks that were detected in 5.3.1, the Inter-beat-interval tachogram was calculated as follows.

$$IBI_i = R_i - R_{i-1} \tag{5.8}$$

Where Peaks are symbolically labeled R to express their correspondence to the R-waves of the ECG, which are commonly used to calculate IBIs, and i is the index value of the peak series. Initially calculated in samples, the IBIs were then translated into milliseconds.

Data Correction

This section is dedicated to the treatment of ectopic beats, arrhythmic events, missing data, and artifacts in the IBI series.

Outlier Rejection. All IBIs that were considered physiologically impossible were marked as outliers. We used the HR-related thresholds HR_{max} and HR_{min} to calculate the shortest (IBI_{min}) and the longest (IBI_{max}) acceptable interval duration. HR_{max} was determined using 5.9 and HR_{min} was set to 45 beats per minute.

$$HR_{max} = \text{Maximum Physiological HR} - \text{Average Subject Age} \tag{5.9}$$

$$IBI_{min} = \frac{60000}{HR_{max}}[ms] \tag{5.10}$$

$$IBI_{max} = \frac{60000}{HR_{min}}[ms] \tag{5.11}$$

The resulting values for the new IBI-related thresholds IBI_{max} and IBI_{min} were rounded to the next full integer and then used to identify outliers in the IBI series. Afterwards, all marked IBI were replaced using linear interpolation.

Ectopic Beat Identification. To identify IBIs that resulted from ectopic beats we considered Kamath's suggestion. Thereby, a heartbeat interval was considered abnormal when the heartbeat interval increased by more than 32.5% decreased by more than 24.5% compared to the previous heartbeat [18]. Equation 5.12 shows the mathematical expression that was used as thresholding method for ectopic beat identification.

$$IBI_{EB} = \{IBI | 0.675 IBI_{n-1} < IBI_n < 1.245 IBI_{n-1}\} \tag{5.12}$$

Artifact Rejection To increase the robustness of the HRV analysis all regions with the presence of ectopic beats or noise for more than 3 seconds were removed entirely [19].

Linear Interpolation. We used linear interpolation to replace outliers and ectopic beats. First, IBIs immediately preceding and after the ectopic beat were marked for replacement. Afterwards, the total time encompassed by all marked IBIs was determined and the number of new beats B, that were inserted to replace them, was computed equation 5.13 for single ectopic beats and equation 5.14 for a sequence of ectopic beats.

$$B = \frac{IBI_i + IBI_{i+1}}{(IBI_{i-1} + IBI_{i+2})/2} \tag{5.13}$$

$$B = \frac{\sum_{j=i}^{i+N} IBI_i}{(IBI_{i-1} + IBI_{i+N+1})/2} \tag{5.14}$$

As B would in general not be an integer value, the computed values were rounded to obtain the nearest integer. After the determination of the number of IBIs to insert was completed, the values of said IBIs were computed using linear interpolation. The basic principle is shown in figure 5.2.

Data Rejection. For short term recordings, such as the ones we conducted in this thesis, it is recommended that at least 80% of a 5 minute segment of data should contain acceptable IBIs [19]. Therefore, recordings that failed this condition were bared from the subsequent feature extraction.

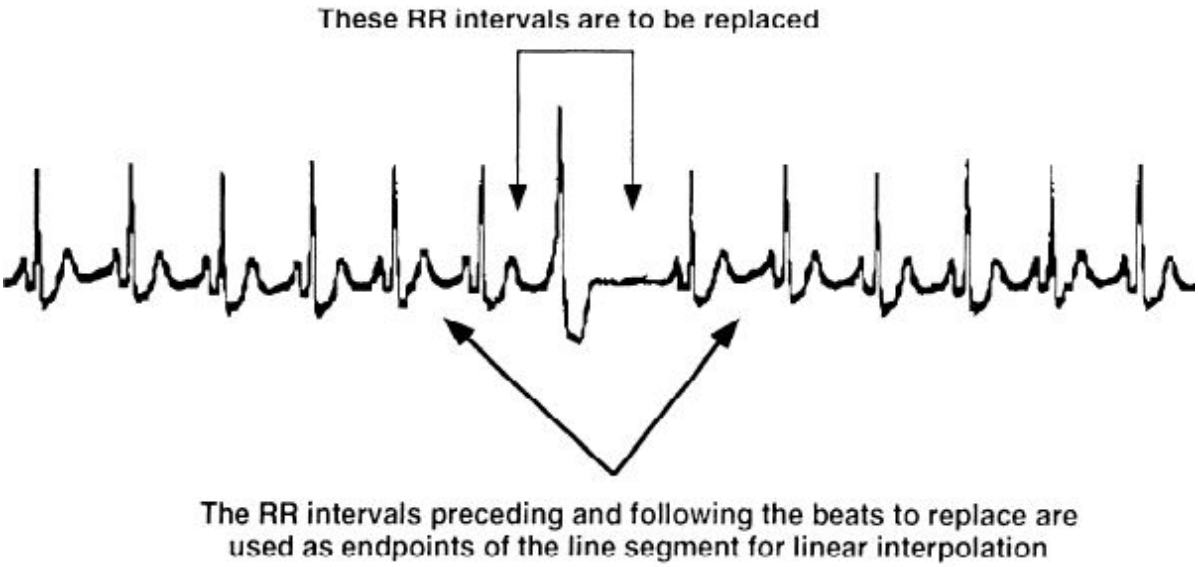


Figure 5.2: A schematic of linear interpolation of ectopic beats on an ECG signal. RR-intervals before and after the ectopic beat are marked for replacement (vertival arrows). RR-intervals before and after the marked segment are used for interpolation (diagonal arrows). source: [20]

Feature Extraction

HRV feature extraction was based entirely on the recommendations given by the Task Force (1996). The selected measures are shown in the tabular below sorted by their domains.

5.3.2 GSR

Pre-Processing

As we were only interested in the tonic component of GSR, pre-processing was focused primarily on artifact removal and noise reduction. The following paragraph covers all measures that were applied in the process of pre-processing.
Lowpass Filtering. A zero-phase fourth-order Butterworth lowpass filter, with a cutoff frequency of 1.0 Hz was implemented to remove high frequency artifacts, caused by electrode movement or electrical noise.
Mean Correction. To account for environmental influences such as temperature and humidity the filtered signal was corrected using the mean GSR value of the baseline

Table 5.1: HRV Feature Selection

Time Domain Measures

Variable	Units	Description
		Statistical Measures
SDNN	*ms*	Standard deviation of all inter beat intervals
RMSSD	*ms*	The square root of the mean of the sum of the squares of differences between adjacent inter beat intervals
SDSD	*ms*	Standard deviation of differences between adjacent inter beat intervals
IBI50		Number of pairs of adjacent inter beat intervals differing by more than 50 ms in the entire recording
pIBI50	%	IBI50 divided by the total number of all inter beat intervals
mIBI	*ms*	Mean of all inter beat intervals
mHR	*bpm*	Mean heart rate of the entire recording
rIBI	*ms*	The difference between the longest and the shortest inter beat interval

Frequency Domain Measures

Variable	Units	Description	Frequency range
		Analysis of short-term recordings (5min)	
Total power	ms^2	The variance of inter beat intervals over the temporal segment	$\leq$ 0.4 Hz
VLF	ms^2	Power in very low frequency range	$\leq$ 0.04 Hz
LF	ms^2	Power in low frequency range	0.04-0.15 Hz
nLF	*n.u.*	LF power in normalized units LF/(Total Power–VLF)·100	
HF	ms^2	Power in high frequency range	0.15-0.4 Hz
nHF	*n.u.*	HF power in normalized units HF/(Total Power–VLF)·100	
LF/HF		Ratio of LF power to HF power	

Table 5.2: GSR Feature Selection

Time Domain Measures

Variable	Units	Description
		Statistical Measures
mGSR	μS	Mean value of the entire recording
rGSR	μS	The difference between the lowest and the highest value
sdGSR	μS	The standard deviation of the entire recording

measurement. The mean corrected data was then subjected to the feature extraction methods.

Feature Extraction

GSR feature extraction was focused on deriving a number of basic statistical time domain measures. The selected measures are shown in the tabular below.

5.3.3 Skin Temperature

Pre-Processing

The pre-processing of the skin temperature signal was identical to the methods applied to the GSR signal. Therefore, we refer to section 5.3.2 at this point.

Feature Extraction

Similar to 5.3.2, feature extraction was focused on deriving a number of basic statistical time domain measures. The selected measures are shown in the tabular below.

Table 5.3: Skin Temperature Feature Selection

Variable	Units	Description
Time Domain Measures		
Statistical Measures		
mST	$^{\circ}C$	Mean temperature of the entire recording
rST	$^{\circ}C$	The difference between the lowest and the highest value
sdST	$^{\circ}C$	The standard deviation of the entire recording

5.4 Data Interpretation

Similar to data analysis, we handled data interpretation using the programming language Python in combination with the PyCharm IDE. In particular, we used the scikit-learn package to implement data and feature preparation methods, as well as parameter optimization in the training process.

5.4.1 Preparation

Formatting. During the feature extraction process, extracted features from one session were separated by signal type and stored in individual Python dictionaries. Therefore, a subroutine was built to extract the features from their individual dictionaries and sort them by the session they originated from. The resulting data points were comprised of all features that were extracted from one recording session (e.g. a 5 minute baseline recording). Afterwards, those data points were converted into a NumPy array with the dimensions $n \times f$, where n was the number of data points (i.e. the sum of all recorded sessions of all the subjects) and f was the number of features that were extracted for each data point.
Generating Training and Test Sets. A common practice in machine learning to prevent a classifier from overfitting is to reserve some of the data for testing and validation purposes. Therefore, the data set was divided into training and test samples. Approximately 25% of data points were reserved for algorithm testing, and the remaining 75% were used for training purposes. This step was implemented using the _train_test_split helper function in the scikit-learn toolbox.
K-fold Crossvalidation is a convenient way to facilitate the evaluation of the training process without having to resort to the test samples. By employing k-fold Crossvalidation

on the training set we were able to compensate for the small sample size of our data set while still reserving the test samples for the final validation.
Feature Preparation. The performance of some machine learning techniques is extremely dependent on data representation. In particular, SVM and Neural Networks benefit from data Normalization (i.e. features are represented as values between 0 and 1). Therefore, we used Python's MinMaxScaler function to translate each feature individually to a range of 0 to 1.

5.4.2 Data Sets

A number of subsets were created to explore classification accuracy for a variety of different scenarios. Therefore, data points were sorted by their session type and recombined to build the subsets shown in 5.4.

5.4.3 Feature Sets

Optimizing feature selection is a key component for the success of machine learning. Therefore, a number of feature sets were created to explore the effects on the classification accuracy of the chosen machine learning algorithms. The individual sets are listed below.

- **Original:** This set was comprised of 21 features from all three physiological signals. Including time domain methods of HRV, GSR, and skin temperature, as well as HRV frequency domain methods.

- **Reduced T:** This set was comprised of 14 features of all three physiological signals. Including only time domain methods of HRV, GSR, and skin temperature.

- **Reduced F:** This set was comprised of 13 features of all three physiological signals. Including only HRV frequency methods, as well as time domain methods of GSR, and skin temperature.

- **Single Sets:** These sets were only comprised of a certain feature group of one physiological measures. Resulting in a set for each of the following: HRV time measures (8 features), HRV frequency measures (7 features), GSR time measures (3 features), and skin temperature time methods (3 features).

Table 5.4: Machine Learning Data Sets

Set Name	**Description**	**Targets**
	Complete Sets	
Original	The set contains all recording sessions. Each session type has an individual label.	6
Rest Vs All	The set contains all recording sessions. All session types, except for the baseline sessions, have the same label.	2
Stress Vs All	The set contains all recording sessions. All session types, except for the stress sessions, have the same label.	2
Emotion Vs All	The set contains all recording sessions. All session types, except for the visual stimulation sessions, have the same label.	2
	Reduced Sets	
Stress Vs Rest	The set contains only stress and baseline sessions. Each session type has an individual label.	2
Stress Vs Stress	The set contains only stress sessions. Each session type has an individual label.	2
Emotion Vs Emotion	The set contains only visual stimulation sessions. Each session type has an individual label.	2

5.4.4 Data Classification

As described in section 1.4.1, we selected the following four machine learning algorithms for classification.

1. **k-NN Classifier**

2. **Random Forest Classifier**

3. **Support Vector Machines**

4. **Neural Networks**

After coding subroutines for each of the above, using their respective implementations in scikit-learn we conducted a gridsearch for algorithm specific hyper-parameters with the goal of performance optimization. Hyper-parameters are parameters that are not directly learned within estimators. In scikit-learn they are passed as arguments to the constructor of the estimator classes (e.g. C, kernel and gamma for SVM). The gridsearch was managed using the GridSearchCV method, which exhaustively generates candidates from a grid of parameter values (param_grid). Tables 5.5 and 5.6 show the parameter grids that were employed in this process. Also, k-fold Crossvalidation (k=3, 5) was applied in the process.

Afterwards, the best estimators for each feature and data set combination were determined by comparing crossvalidation scores. Lastly, the final validation was conducted using the best estimators on the reserved test data sets.

Table 5.5: Gridsearch Parameters Part 1

Parameter	Description	Values
	k-NN Classifier	
n_neighbors	The number of neighbors that is considered.	1-12
	Support Vector Classifier	
kernel	Specifies the kernel type to be used in the algorithm.	rbf
C	Regularization parameter. The strength of the regularization is inversely proportional to C.	0.001, 0.01, 0.1, 1, 10, 100, 150, 200
gamma	The kernel coefficient.	0.0001, 0.001, 0.01, 0.1, 1, 10, 50, 100

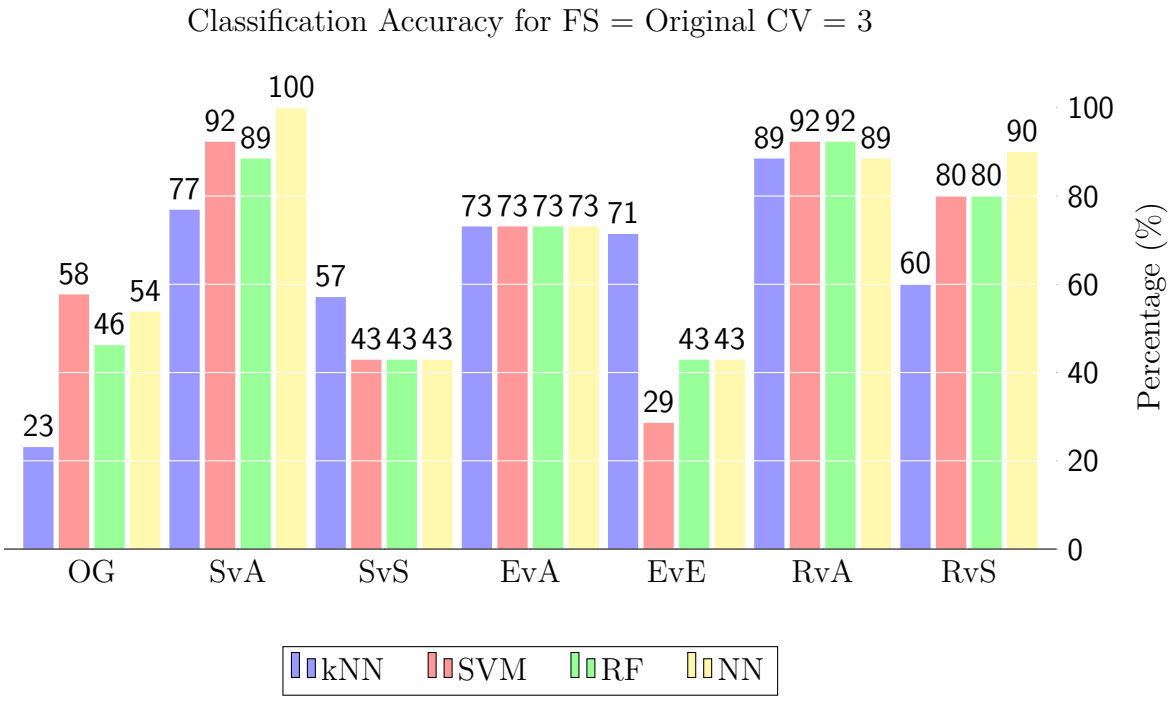
Table 5.6: Gridsearch Parameters Part 2

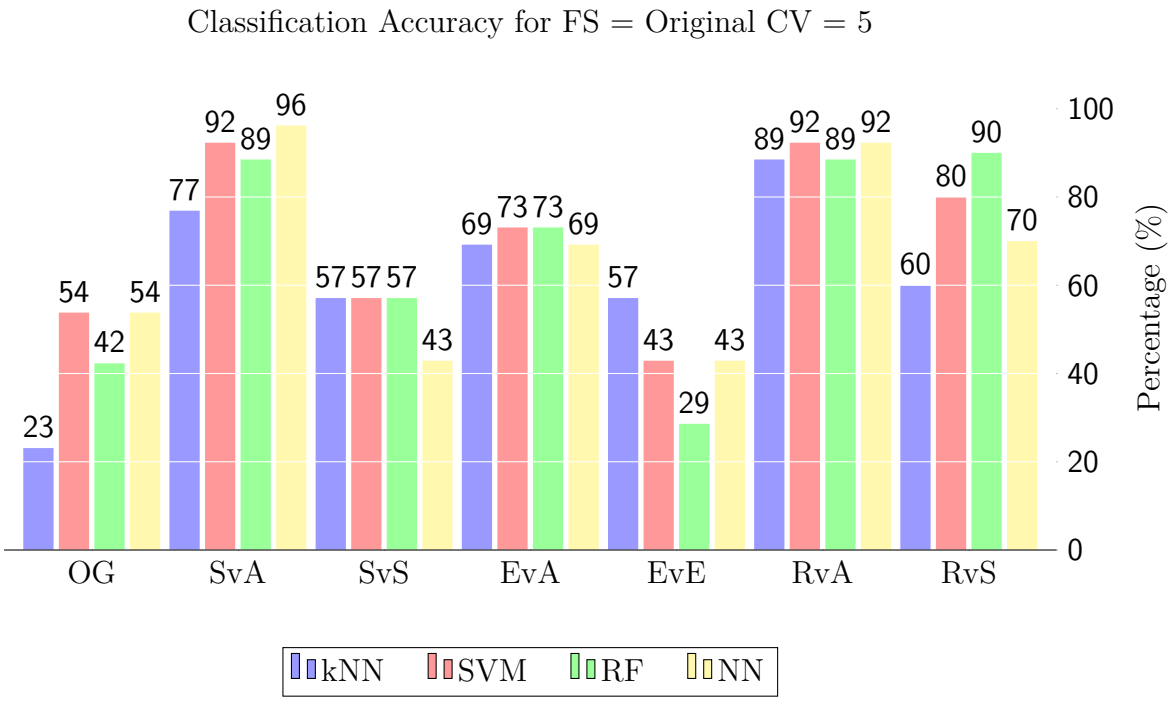
Parameter	Description	Values
	Random Forest Classifier	
n_estimators	The number of trees in the forest	5, 25, 50, 100, 250, 500, 1000
max_depth	The maximum depth of the tree.	1-5
max_features	The number of features to consider when looking for the best split. Because we use float values int(max_features $\cdot$ n_features) features are considered at each split.	0.10, 0.25, 0.5, 0.75, 1.0
	Neural Networks	
solver	The solver for weight optimization.	lbfgs
activation	The activation function for the hidden layer.	tanh, relu
hidden_layer_size	The number of neurons in each layer. We used 1-3 layers, with identical amounts of neurons, determined by int(max_features $\cdot$ hidden_layer_size).	0.5, 1, 2
alpha	L2 penalty (regularization term) parameter.	0.00001, 0.0001, 0.0005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.5, 1
learning_rate	The learning rate is used to control the step-size in updating the weights.	0.001, 0.01, 0.1

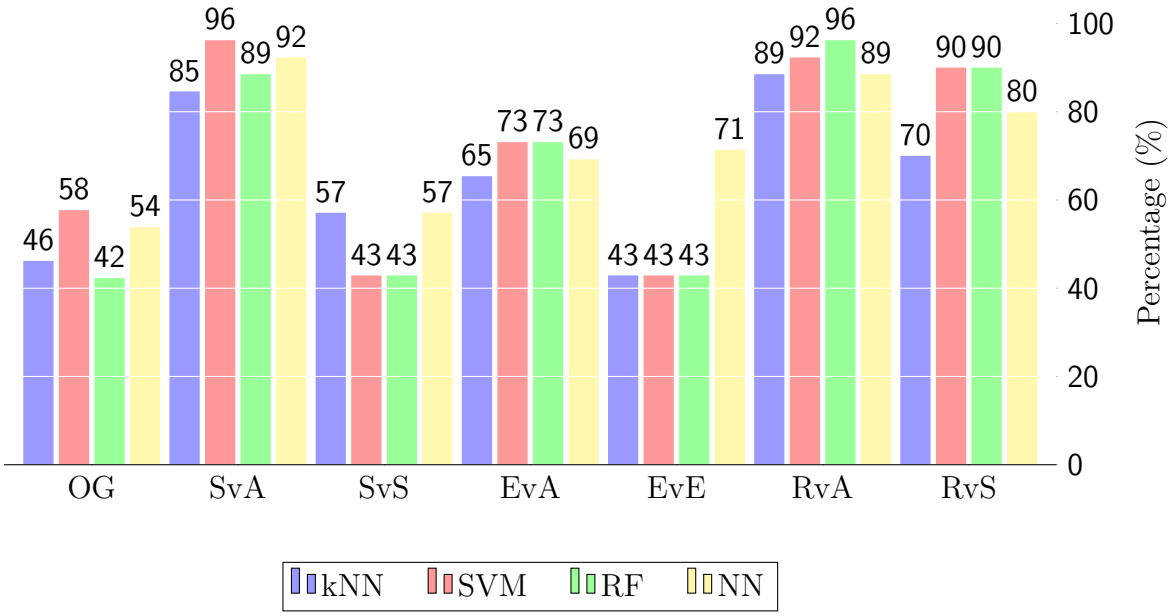
6 Results

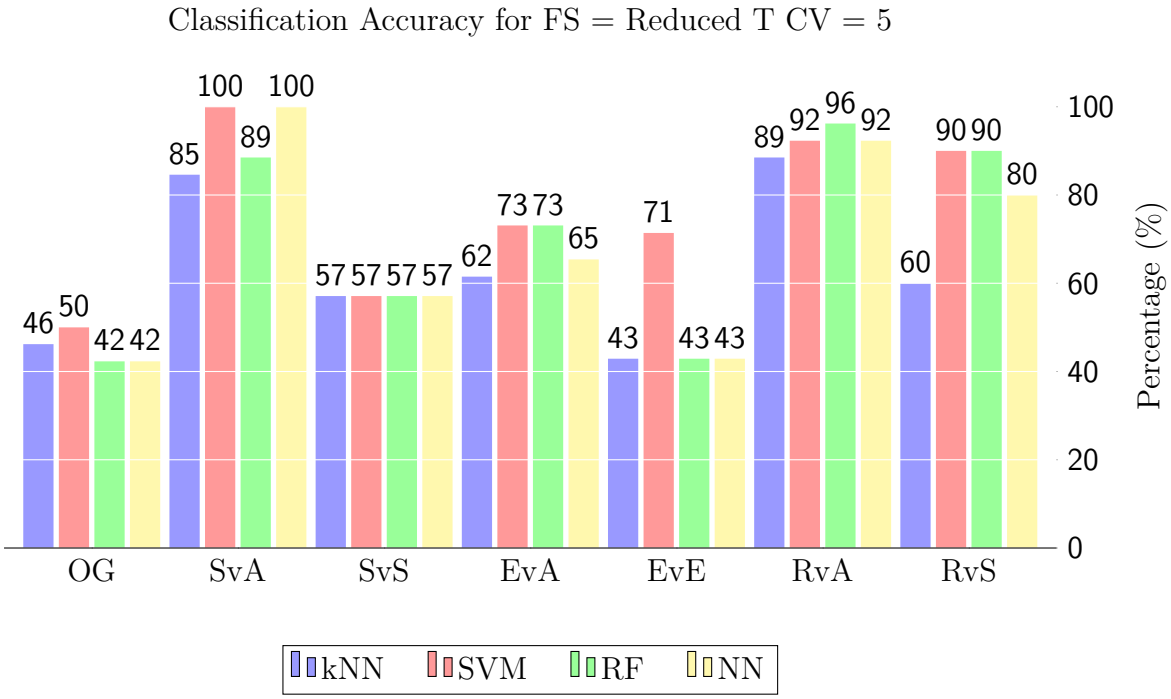
In this chapter we present the results of the gridsearch we conducted in the process of parameter optimization. Using accuracy as a measure for the classification quality we evaluated the effects of feature and data selection on the generalization ability of our machine learning algorithms.
Classification accuracy has been calculated from classifier validation (i.e. using the trained estimator to make predictions on the reserved test data).

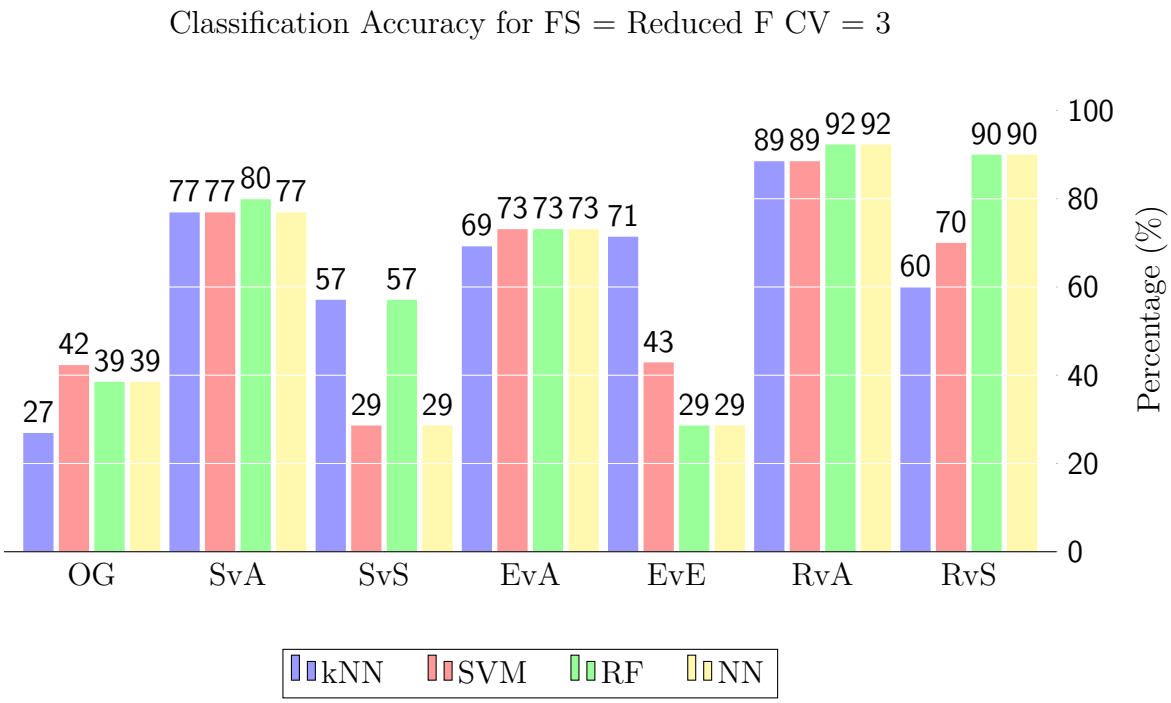
The results of the validation process are shown in the graphs below. We used a series of bar plots to visualize the accuracies we achieved with each classifier for each data set based on the used feature set. Plots are distinguished by feature selection and Crossvalidation. Data sets, on the x-axis, are plotted against accuracy, on the y-axis. The bars are color coded to indicate machine learning algorithms.



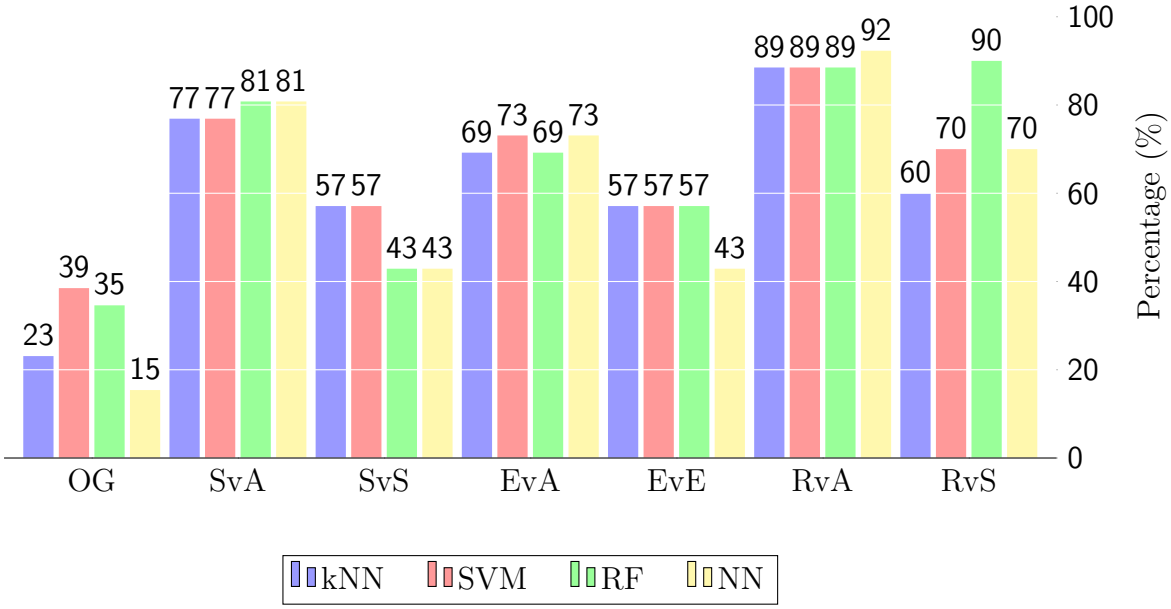

Classification Accuracy for FS = Original CV = 5
23
54
42
54
77
92
89
96
57
57
57
43
69
73
73
69
57
43
29
43
89
92
89
92
60
80
90
70
OG
SvA
SvS
EvA
EvE
RvA
RvS
0
20
40
60
80
100
Percentage (%)
kNN
SVM
RF
NN


Classification Accuracy for FS = Reduced T CV = 3
46
58
42
54
85
96
89
92
57
43
43
57
65
73
73
69
43
43
43
71
89
92
96
89
70
90
90
80
OG
SvA
SvS
EvA
EvE
RvA
RvS
0
20
40
60
80
100
Percentage (%)
kNN
SVM
RF
NN

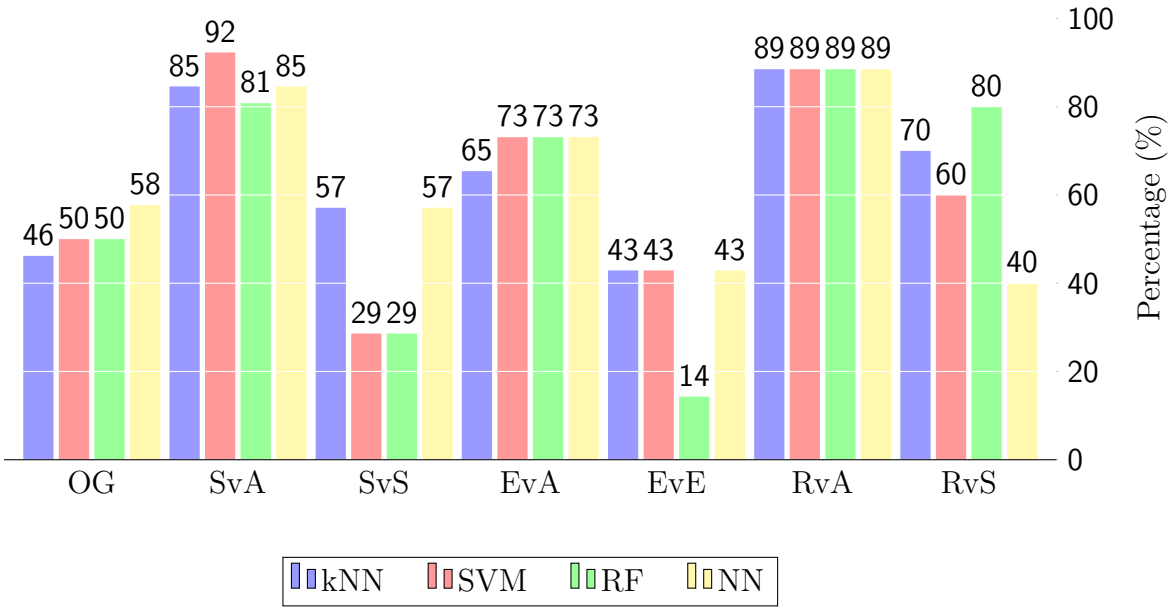

Classification Accuracy for FS = Reduced T CV = 5
46 50 42 42
85 100 89 100
57 57 57 57
62 73 73 65
43 71 43 43
89 92 96 92
60 90 90 80
OG
SvA
SvS
EvA
EvE
RvA
RvS
0 20 40 60 80 100
Percentage (%)
kNN
SVM
RF
NN


Classification Accuracy for FS = Reduced F CV = 3
27 42 39 39
77 77 80 77
57 29 57 29
69 73 73 73 71 43 29 29
89 89 92 92
60 70 90 90
OG
SvA
SvS
EvA
EvE
RvA
RvS
0 20 40 60 80 100
Percentage (%)
kNN
SVM
RF
NN

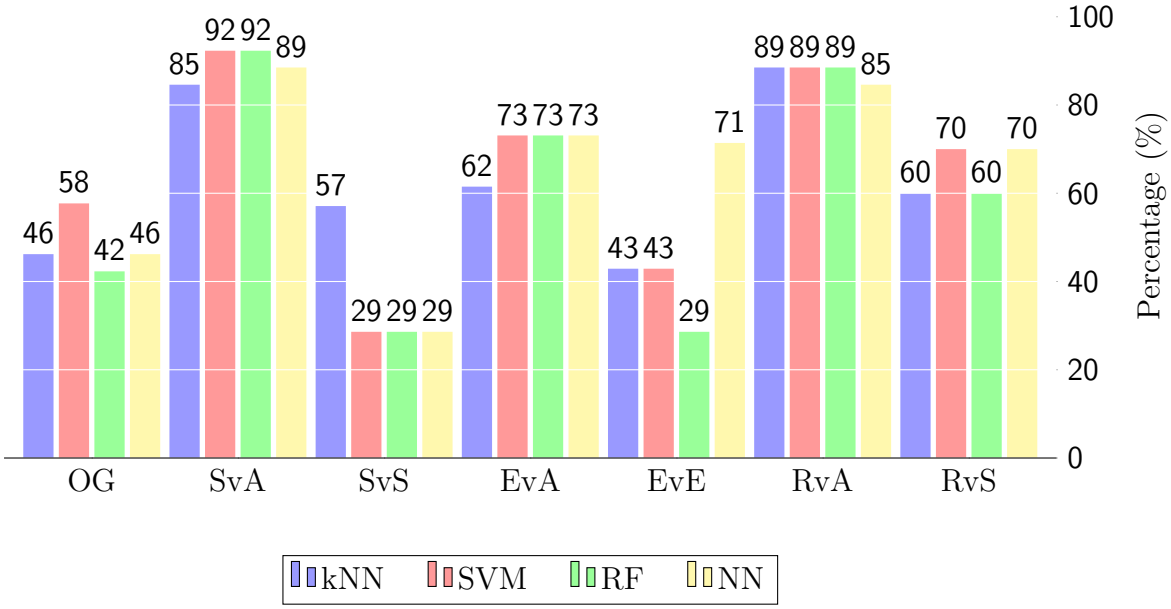
Classification Accuracy for FS = Reduced F CV = 5



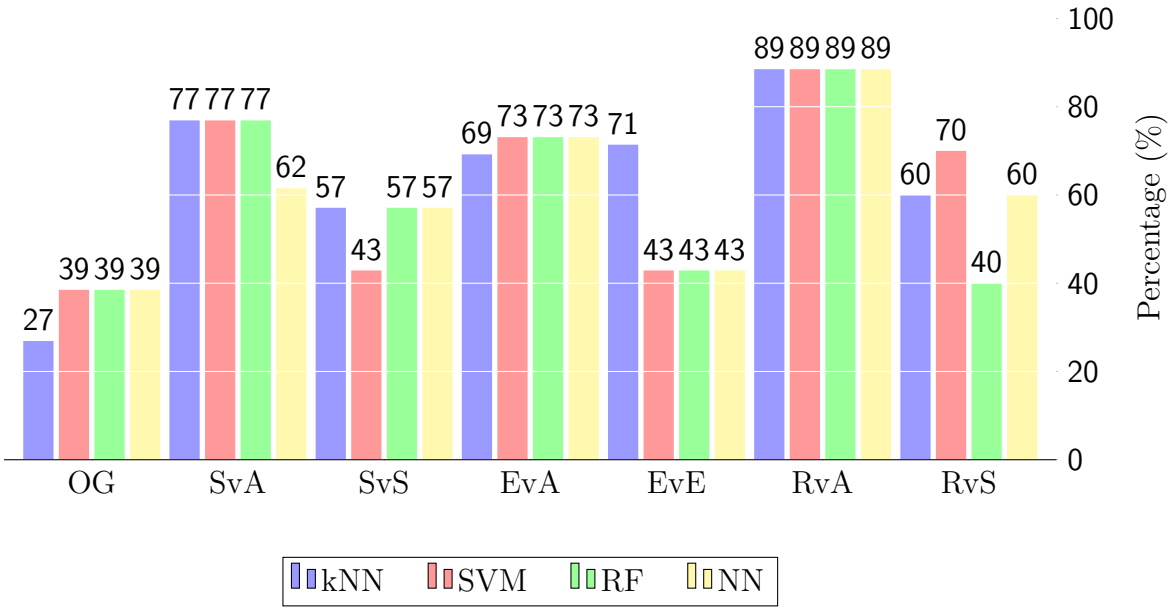
Classification Accuracy for FS = Single HRV time CV = 3



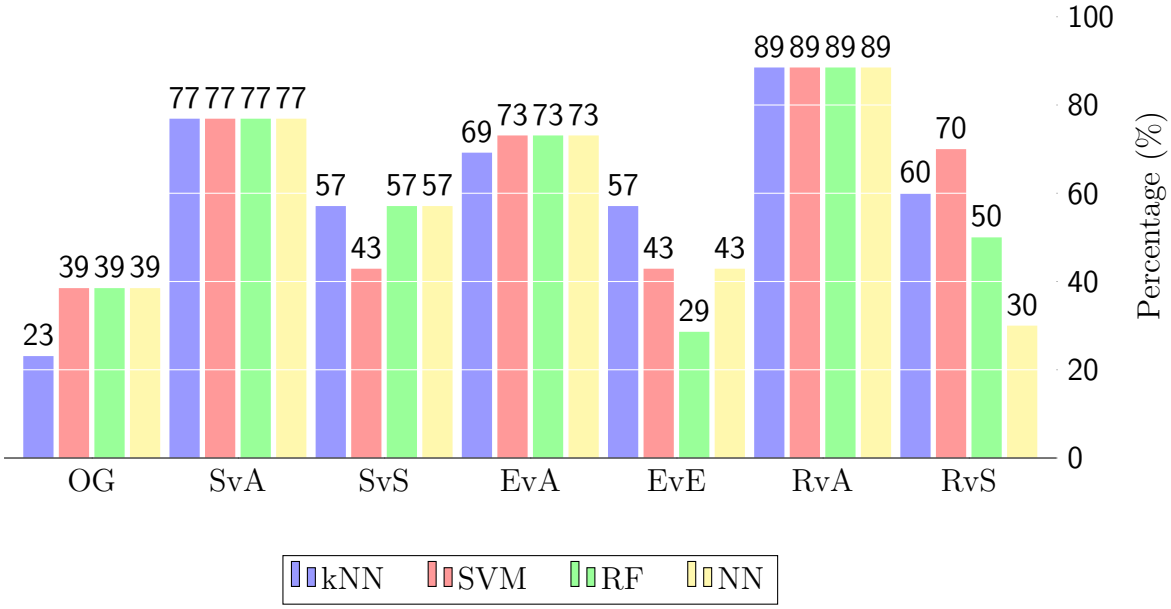
Classification Accuracy for FS = Single HRV time CV = 5



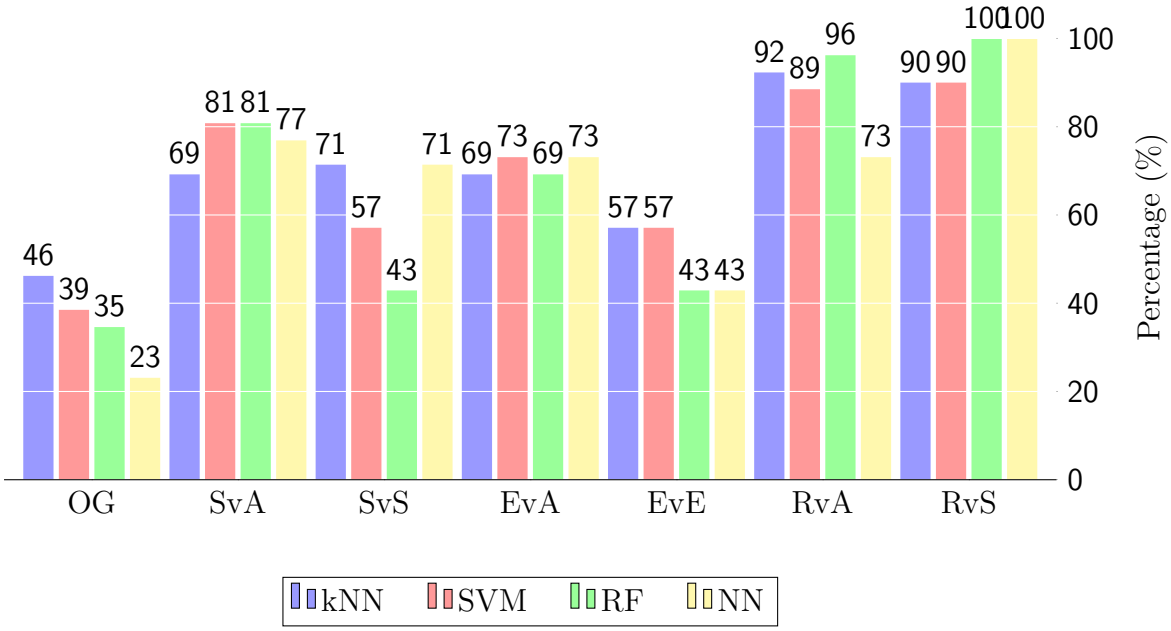
Classification Accuracy for FS = Single HRV freq CV = 3

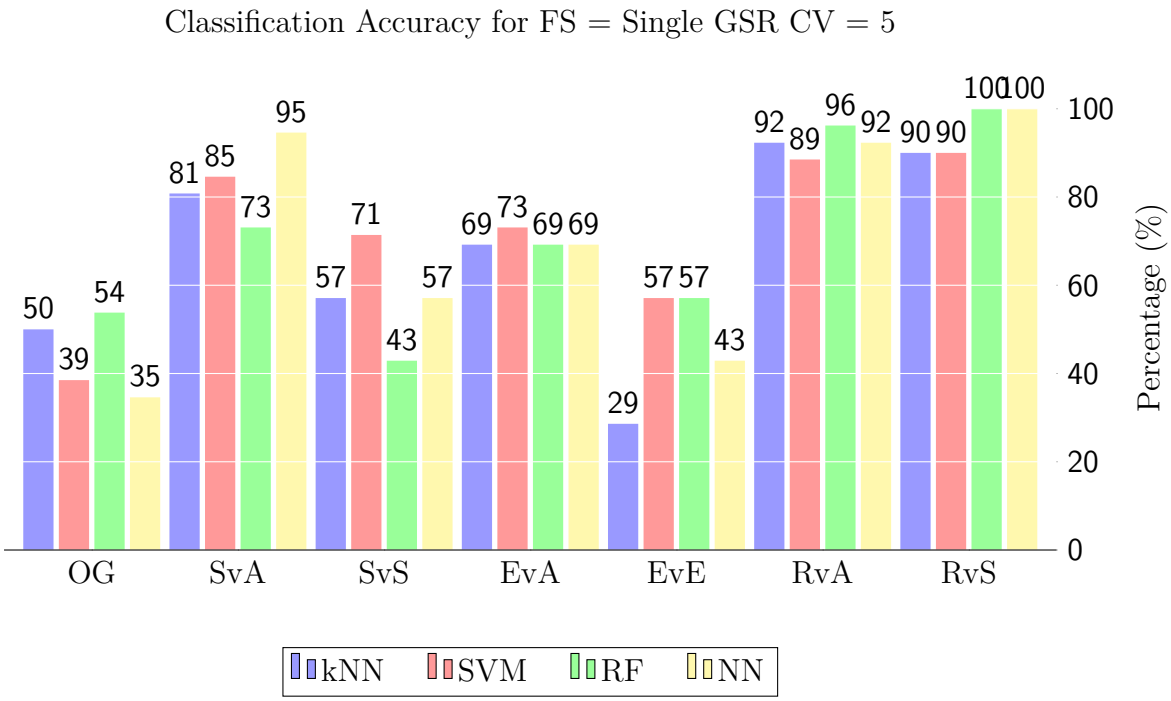


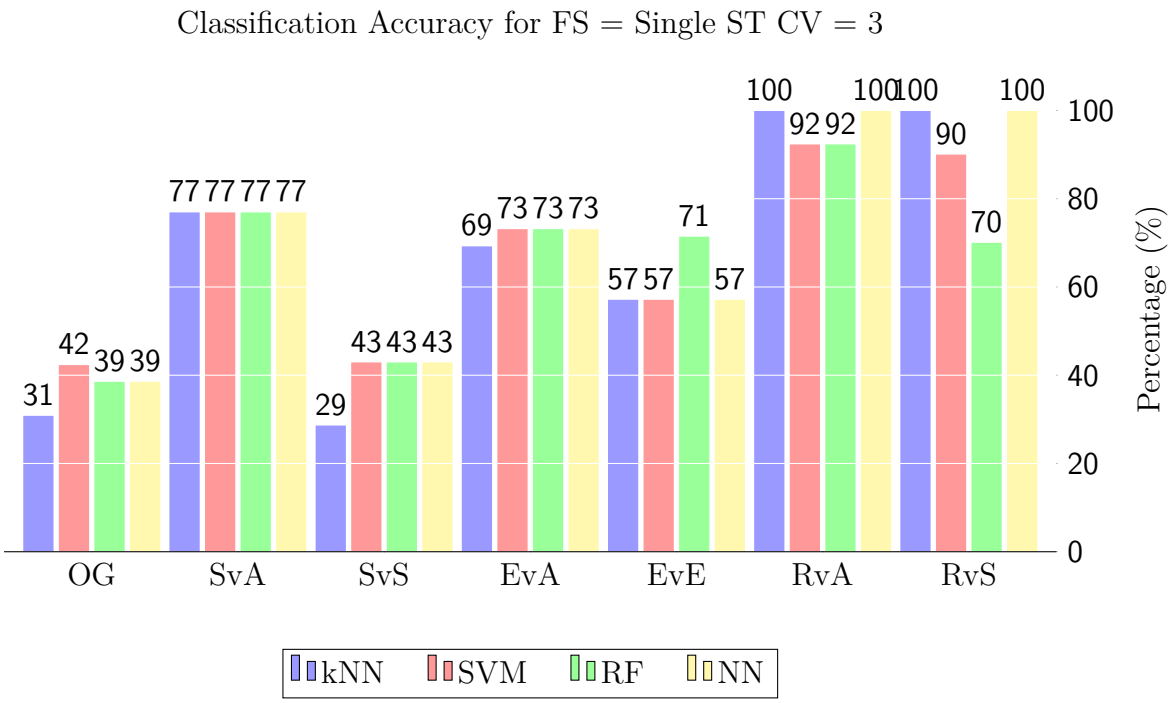
Classification Accuracy for FS = Single HRV freq CV = 5

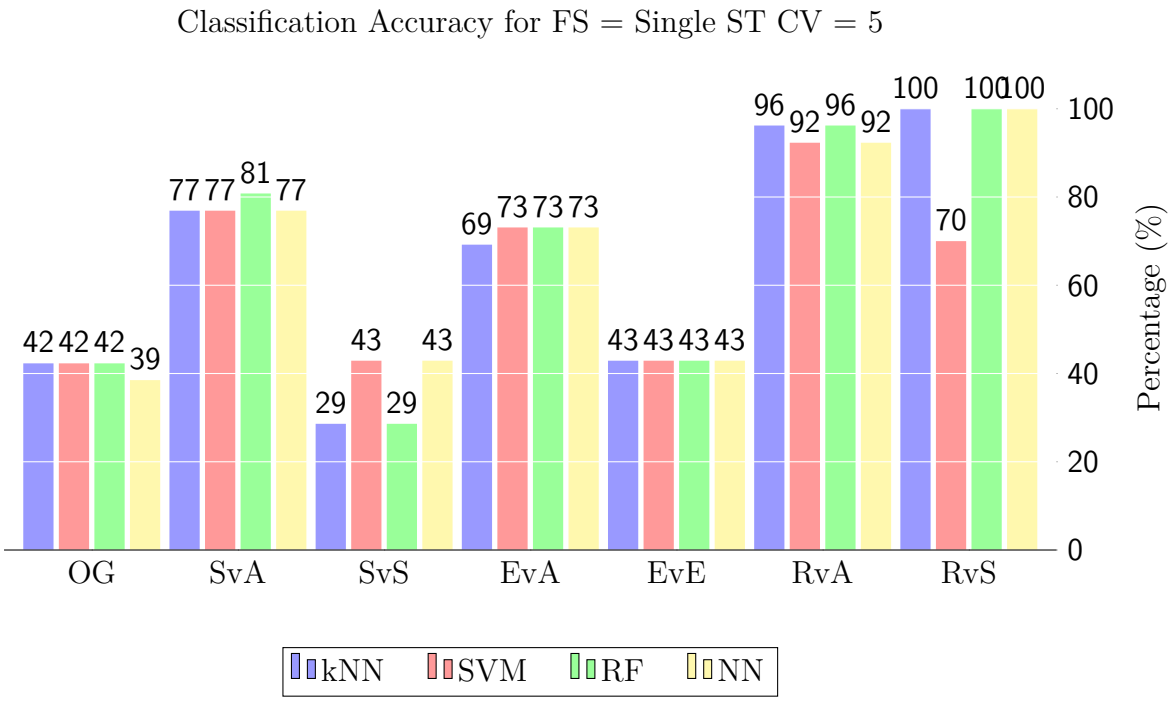


Classification Accuracy for FS = Single GSR CV = 3




Classification Accuracy for FS = Single GSR CV = 5
50 39 54 35
81 85 73 95
57 71 43 57
69 73 69 69
29 57 57 43
92 89 96 92
90 90 100 100
OG SvA SvS EvA EvE RvA RvS
0 20 40 60 80 100
Percentage (%)
kNN SVM RF NN


Classification Accuracy for FS = Single ST CV = 3
31 42 39 39
77 77 77 77
29 43 43 43
69 73 73 73
57 57 71 57
100 92 92 100
100 90 70 100
OG SvA SvS EvA EvE RvA RvS
0 20 40 60 80 100
Percentage (%)
kNN SVM RF NN



However, in this form the results have rather low significance, as the sheer amount of plots hinders proper evaluation. Therefore, we constructed the following tables to improve data illustration. We listed the average accuracy for each classifier based on either feature selection (see 6.1) or data set (see 6.2) to providing insight on their respective effects on individual algorithm performances.

Table 6.1: Feature Based Performance

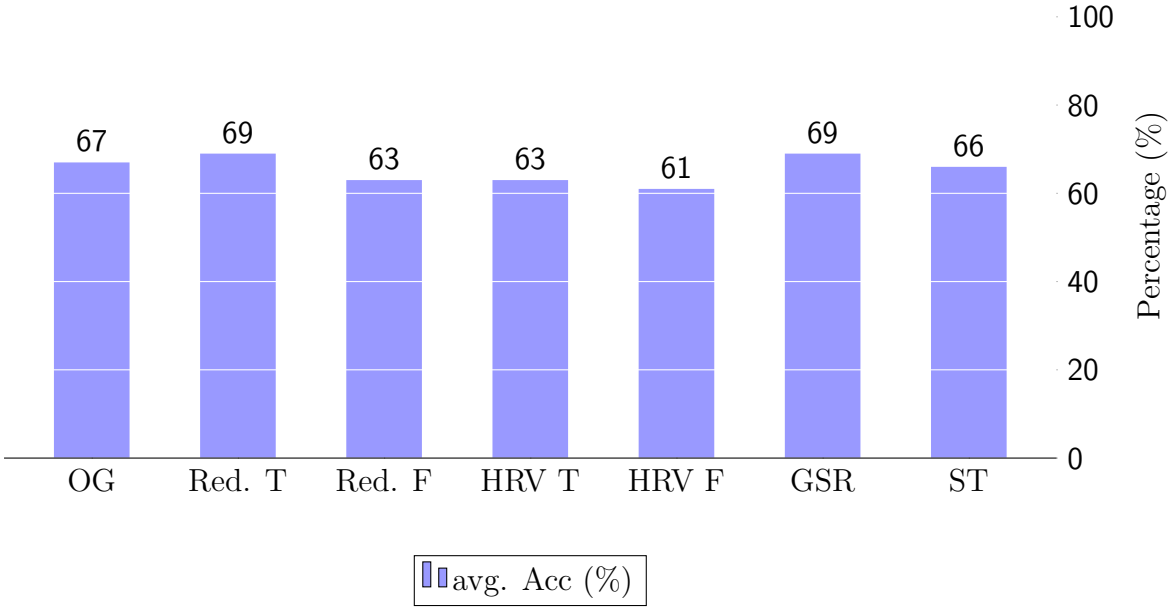
Algorithm	**Accuracy (%)**	**Algorithm**	**Accuracy (%)**
Original		Single HRV Freq.	
kNN	63	kNN	63
SVM	68	SVM	62
RT	68	RT	59
NN	69	NN	59
Reduced T		Single GSR	
kNN	64	kNN	69
SVM	73	SVM	71
RT	69	RT	69
NN	71	NN	68
Reduced F		Single ST	
kNN	63	kNN	66
SVM	63	SVM	65
RT	66	RT	66
NN	60	NN	68
Single HRV Time		Averages	
kNN	64	kNN	65
SVM	64	SVM	67
RT	59	RT	65
NN	65	NN	66

Table 6.2: Data Based Performance

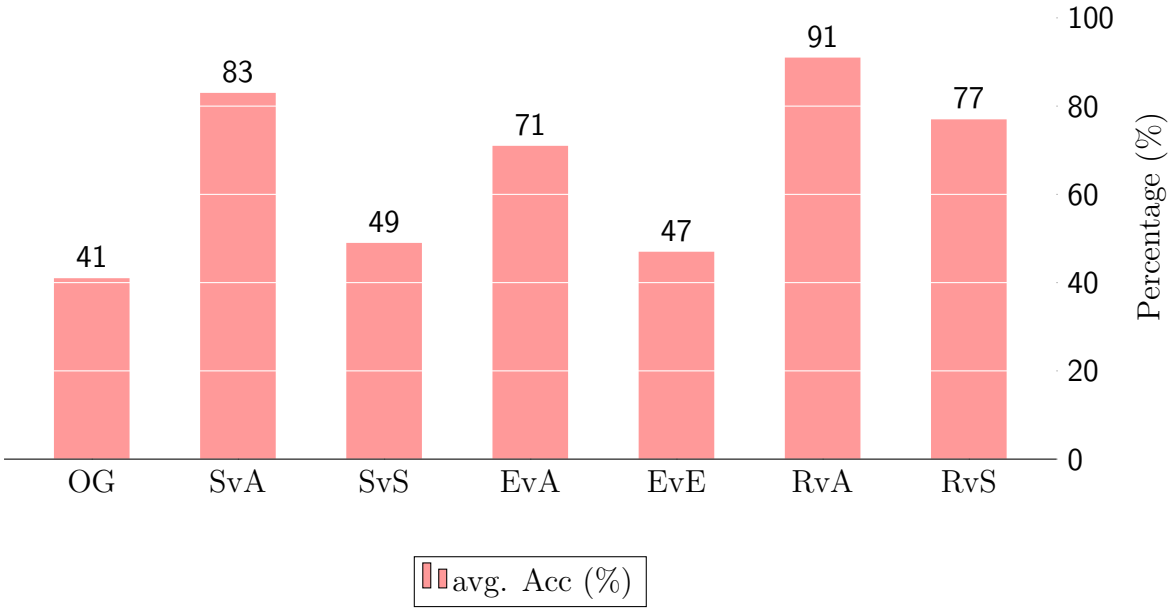
Algorithm	**Accuracy (%)**	**Algorithm**	**Accuracy (%)**
Original		SvA	
kNN	36	kNN	79
SVM	46	SVM	85
RT	42	RT	81
NN	41	NN	85
SvS		EvA	
kNN	54	kNN	68
SVM	46	SVM	73
RT	45	RT	72
NN	49	NN	72
EvE		RvA	
kNN	53	kNN	91
SVM	48	SVM	90
RT	41	RT	92
NN	47	NN	90
RvS		Averages	
kNN	71	kNN	65
SVM	78	SVM	67
RT	81	RT	65
NN	77	NN	66

Lastly, we calculated classification accuracies for individual data sets (see 6), as well as individual feature sets (see 6) using the collective performance average of all four algorithms.

Average Classification Accuracy Feature Based



Average Classification Accuracy Data Based



7 Discussion

7.1 Answering Research Questions

7.1.1 Research Question 1

Is it possible to get real-time access to the data we record with the Empatica E4?

Based upon the findings of the pilot experiment we conducted in section 4.1, we are able to answer in the affirmative. Using the MATLAB TCP client we built in the scope of this project we were able to derive raw signal data from the Empatica E4, using a wireless Bluetooth connection, at the same rate it was recorded.

7.1.2 Research Question 2

Is it possible to detect and distinguish different degrees of mental workload, as well as two emotional states, such as pleasant and unpleasant, using common machine learning techniques on three physiological signals that were recorded with the Empatica E4 in an experimental setting?

With the methods we described in 5.4 we were able to classify a total of 6 different emotional states, with accuracies as high as 58% for individual classifiers, and an average of 44 % for all machine learning algorithms. These values, were derived from individual and collective algorithm performances using the original feature set, with a total of 21 features that were extracted from HRV, GSR and skin temperature.
Further, we were able to achieve even higher performances for binary classification tasks, aiming to identify specific emotional states. Considering the average performances of all four algorithms we achieved accuracies of 89% for stress sessions, 72% for visual stimulation sessions, and 91% for resting intervals using again all 21 features.
Addressing the distinction of similar emotional states, we assessed the performances of all four classifiers using data sets, only comprised of session from either visual stimulation or stress sessions. Considering again the full feature set we achieved performances of 77% for resting and stress conditions, and 45% for both medium and high stress levels, as well as emotional states.

Lastly, we report feature dependent accuracies of 61-69% for all of the abovementioned classification tasks. These values were computed using the average performances of all four classifiers on the respective feature sets.
Considering these results, we conclude that our methods are more sensitive to the presence or absence of a certain state opposed to small or gradient changes. But much like Byrne et al. (1996) we believe this type of "coarse" state assessment not necessarily to be a problem. The application of such a binary assessment of workload or emotion may still provide significant information to adaptive logic (e.g. on-task versus off-task) [2]. Therefore, we again answer in the affirmative and believe in the potential of our system for the neuroergonomic assessment in adaptive workplaces.

7.1.3 Research Question 3

Which algorithm is best suited for the classification task introduced in RQ2?

We found that overall SVM with 5-fold Crossvalidation performed best when used in combination the reduced time feature set. However, when compared over all feature sets all four algorithms achieve similar results ranging from 65 to 67 %.

7.2 Success and Limitations of Neuroergonomic Assessment

Although, we achieved accuracies well above chance for the classification of 6 different emotions using our selected features, and reached even higher performances on the binary classification of subsets using a variety of feature combination, we still believe that there is room for improvement. Therefore, in the following section, we will discuss the limitations of the present and the possibilities of future work.

7.2.1 Limitations

Signal Quality. In the sense of the concept of "Garbage In Garbage Out" the recent work focused heavily on the quality of data extraction and processing methods. The overall morphology and quality of the signal we derived, using the E4 wristband, were adequate and similar to other commercial devices. Further, we were not able to identify any drop-off in quality caused by the derivation from the standard measurement locations of PPG and GSR. We did however notice disproportional artifact levels in the BVP signal during break intervals of our experiment, in which the restrictions to the participants

movement were lifted. This suggests the susceptibility of wrist based signal detection to motion artifacts caused by muscle activity and subsequent sensor displacement. Even though these implications cast an unfavorable light on the E4, we could argue that this is an inherent problem to all optical sensors. However, at this point and without further investigation we are not able to determine the effects on future applications.

Crossvalidation. We have already touched upon the use of k-fold Crossvalidation as a way to increase the performance of machine learning algorithms. However, our results do not clearly indicate the benefits of different degrees of Crossvalidation. Again, we propose that the limited number of data points to be a possible reason for this phenomenon. Thereby, using an excessive degree of k-fold Crossvalidation, which causes the individual validation sets to shrink and ultimately a distortion in the results, will lead to sub par or worse performances. Therefore, we would expect to observe further improvement by extending the data base and reassess k-fold parameters. One possibility could be to use the Leave One Out method. It is a rather extreme form of k-fold Crossvalidation, in which each sample is used once as a test set (singleton) while the remaining form the training set. We applied this method in the earlier stages of development, and very quickly ruled it out due to its extremely high computational cost. To conclude, even if the benefits of k-fold Crossvalidation were not immediately noticeable in our experiment its general usefulness and potential in future applications should be considered.

Evaluation Metric. In the scope of this thesis we used accuracy as a primary metric for algorithm evaluation. Although this is common practice, accuracy is known to be deceptive in some cases. For example, training learning algorithms on asymmetrical data sets. As most of our binary classification tasks featured a very asymmetrical data base (i.e. one of the emotional states was heavily favored), we faced some complications in the learning process. Especially during the classification of visual stimulation sessions in the complete data set. We achieved very similar performances for all algorithms across all feature sets, which lead to doubts in the validity of the results. Therefore, we employed additional metrics such as the Confusion Matrix, which indicates the number of correct and incorrect decisions for each of the target labels. Applying this method revealed faulty decision patterns for most of the algorithms on the above mentioned data set, which suggested some issues in the learning process. We believe that the small sample size in combination with the strong imbalance of the data set are the most likely reasons for these results and that the best approach would be to conduct further experiments, creating a larger data set with more equally distributed states.

Sample Size. Sample size, or the size of the data set, is known to be a great influence to the success of machine learning algorithms. Considering the algorithms that were selected for this project, we were able to witness the effects of sample size first hand. For example we noticed that algorithms such as Random Forest, Support Vector Machines, and Neural Networks, which are all known to be better suited for medium to large data

sets, significantly underperformed on most of the smaller sub sets (i.e. SvS, EvE) when compared to the larger subsets (i.e. SvA, EvA, RvA). Also, the k-NN classifier, which is commonly known to perform well even on thinner data sets, performed exceptionally compared to the others on distinguishing among similar emotional states (i.e. SvS, EvE). Considering these results, we believe that the acquisition of further data could greatly improve the performance of some of the more complex learning algorithms and our system overall.

Emotion Elicitation. According to our initial approach we used visual stimulation to elicit two emotional states in our subjects: pleasant and unpleasant. As described in 4.2.3 we presented a selected set of images, taken from the IAPS, for each emotion. Image selection was based on rating according to a two dimensional model (arousal and valence). In addition, subjective ratings (which used the same metric) on all presented images were gathered for all subjects directly after the visual stimulation sessions. After careful evaluation of the subjective ratings (supported by our observations during the experiment) we detected a number of discrepancies. Meaning that the emotion participants perceived from a certain picture would deviate from the one we intended to display. This misconception raised our concern due to its inevitably effects on the quality of the features we derived from the respective sessions. But, because this phenomenon was limited to a minuscule percentage of data points and the fact that our subject group was already very limited in size, we hesitated to filter out the flawed sessions. It is not clear to which extent the results of the respective classification tasks (i.e. EvA, EvE) were affected by this decision. On the other hand, it could be argued that these findings were coincidental and would have been rendered insignificant in a larger group of people. Nevertheless, we propose a more elaborate emotion elicitation paradigm, which firstly separates neutral and pleasant stimuli and secondly features stimuli that are more representative of the respective emotions, for future iterations.

Qualification Criteria As described in section 4.2.2 we included a total of 14 healthy subjects in our experiment. All of which reported to be healthy and willing to partake in the experiment. In retrospective we identified the following issues with the subject admission process.

- Recruited from the student body of HTW Saar and SNNU, most of our subjects came from a scientific background and were already familiar with the laboratory and experimental procedures. In addition, some of the participants had already worked with the IAPS. We can only assume the effects this familiarity had on the emotion elicitation capability of our visual stimulation tasks.

- Although all participants gave their confirmation, some stated afterwards that they either were not feeling well, suffered from bad mood, or a mild headache but still wanted to partake. Again we can not completely estimate the influence of this on the measurement, or rule out that there is any influence at all

Classification Interval. As we were dedicated to follow the recommendations of the Task Force of The European Society of Cardiology and The North American Society of Pacing and Electrophysiology (1996) we used HRV features that were extracted from 5 minute recordings in our emotion classification process. However, it is also stated that the necessary interval length lies only between 1 min for HF and 2 min for LF, as the recording interval should be at least 10 times the length of the wavelength of the lowest frequency component. Considering real life settings, the reduction in classification time we could provide with this method, could greatly benefit the responsiveness and the overall accuracy of adaptive automation. Furthermore, we argue that due to the E4's susceptibility to motion artifacts the classification on 5 minute intervals is far too impractical for real world applications. However, the effects on HRV frequency features should be evaluated first before deploying such a method.

7.2.2 Future Work

In this section we present a list of possible additions to the recent work, that were derived during the evaluation process and could be the addressed in subsequent projects.

Large Scale Data Collection. Over the course of this thesis we encountered a number of issues with the machine learning process (see 7.2.1). Many of which could be resolved by significantly extending the data base. Therefore, we would like to propose a subsequent large scale experiment, to collect proof our hypothesis and for the further improvement of the system at hand. Next to a larger subject group, we identify the refinement of the admission criteria as well as the emotion elicitation paradigm as key components for such an endeavor.

HMI Focused Application. Unfortunately, we were unable to realize an experimental setting within an actual collaborative work environment in the given time frame. Therefore, we would like to see a continuation of our work which is focused on the development of adaptive logic based on the application of the current system in an collaborative work environment.

Motion Compensation. Our ultimate goal is the application of the recent system in a real life environment. Because of this we need to address problems that are likely to arise once we step outside of laboratory conditions. We consider motion artifacts to be one of the major obstacles to system implementation. Therefore we propose the investigation of motion based signal processing strategies, which attempt to adapt signal detection and analysis based on acceleration data derived from the E4's accelerations sensors.

Weighted Fusion Strategy To our surprise single signal feature sets such as GSR performed well on many of the classification tasks. Inspired by the work of Wei et al. (2018) we propose to investigate the effects of Weighted Fusion Strategy on the recent work.

Linear Discriminant Analysis Considering the work of Maaoui and Pruski (2010) who achieved a classification accuracy of 92% over 6 emotions by using SVM and Fisher Linear Discriminant analysis, in respect to the results of the present thesis, we were pondering the implications of adapting this strategy. However, due to the limited time frame we were unable to do so. Therefore, we would like to conduct further research on the matter in the future.

8 Conclusion

Our work was guided by our aspirations to facilitate neuroergonomic assessment in collaborative workplaces. To reach this goal, we took on the task of developing a truly unobtrusive monitoring system by employing the Empatica E4 to provide wireless acquisition of physiological data. We established a data base for machine learning by conducting a series of experiments, which were specifically designed to elicit authentic emotion, using the newly developed system in a lab environment. We implemented state-of-the-art signal processing and analysis to derive meaningful features from three different physiological channels (BVP, GSR, and skin temperature) and eventually utilized them in our investigation on suitable learning algorithm for emotion classification.
All things considered, we were pleased by the outcome and the system's overall performance under the given circumstances. We were able to deliver proof of concept for the system we developed and facilitate the neuroergonomic assessment in binary classification tasks. However, we acknowledge the fact that there is room for improvement and further refinement is still in order to guarantee the success of our system in real life settings. We consider both, the extension of the data base, as well as the investigation of more elaborate emotion elicitation methods, the top priorities for subsequent projects.

IAPS selected pictur list motive names

List of Figures

List of Tables

Bibliography

[1] R. Parasuraman. Neuroergonomics: Research and practice. *Theoretical Issues in Ergonomics Science*, 2003.

[2] E. A. Byrne and R. Parasuraman. Psychophyiology and adaptive automation. *Biological Psychology*, 1996.

[3] R. K. Mehta and R. Parasuraman. Neuroergonomics: a review of applications to physical and cognitive work. *frontiers in Human Neuroscience*, 2013.

[4] S. Clarke and C. L. Cooper. *Managing the Risk of Workplace Stress. Health and safety hazards.* Routledge. London and New York, 2004.

[5] R. Parasuraman and Rizzo M. *Neuroergonomics*, volume 3 of *Human-Technology Interaction Series*. Oxford University Press, 2008. ISBN 0195368657, 9780195368659.

[6] J. Allan. Photoplethysmography and its application in clinical physiological measurement. *Physiological Measurement*, 2007.

[7] E. Peper, R. Harvey, I. Lin, H. Tylova, and D. Moss. Is there more to blood volume pulse than heart rate variability, respiratory sinus arrhythmia, and cardiorespiratory synchrony? *Biofeedback*, 2007.

[8] Task Force of The European Society of Cardiology, The North American Society of Pacing, and Electrophysiology. Heart rate variability. standards of measurement, physiological interpretation, and clinical use. *European Heart Journal*, 1996.

[9] X. Wu, V. Kumar, J. R. Quinlan, J. Gosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, P. S. Yu, Z.-H. Zhou, M. Steinbach, D. J. Hand, and D. Steinberg. Top 10 algorithms in data mining. *Knowl Inf Syst*, 2008.

[10] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 2011.

[11] D. C. Rubin and J. M. Talarico. A comparison of dimensional models of emotion:

Evidence fromemotions, prototypical events, autobiographical memories, andwords. *Memory*, 2009.

[12] *E4 wristband from empatica. User's manual.* Empatica, Via Stendhal 36, 20144 Milano (MI). URL www.empatica.com.

[13] E4 streaming server for windows. reference guide, May 2018. URL https://developer.empatica.com/windows-streaming-server.html.

[14] P.J. Lang, M.M. Bradley, and B.N. Cuthbert. International affective picture system (iaps): Affective ratings of pictures and instruction manual. Technical Report A-8, University of Florida, Gainesville, FL., 2008.

[15] S. Jern, M. Pillhall, C. Jern, and S.G. Carlsson. Short-term reproducibility of a mental arithmetic stress test. *Clinical Science*, 1991.

[16] M. Mitrushina, K.B. Boone, J. Razani, and L.F. D'Elia. *Handbook of Normative Data for Neuropsychological Assessment.* Oxford University Press, USA, 2005. ISBN 9780195169300. URL https://books.google.de/books?id=Ygndi8UlmxkC.

[17] M. Elgendi, I. Norton, M. Brearley, D. Abbott, and D. Schuurmans. Systolic peak detection in acceleration photoplethysmograms measured from emergency responders in tropical conditions. *PLoS ONE*, 2013.

[18] A. Choi and H. Shin. Quantative analysis of the effect of an extopic beat on the heart rate variability in the resting condition. *Frontiers in Physiology*, 2016.

[19] G.D. Clifford. *Signal Processing Methods for Heart Rate Variability.* PhD thesis, St. Cross College, 2002.

[20] N. Lippman, K. M. Stein, and B.B. Lerman. Comparison of methods for removal of extopy in measurement of heart rate variablility. *American Journal of Physiology*, 1994.