Master Thesis on Intelligent Interactive Systems

Universitat Pompeu Fabra

Selecting Features of Breathing & Stability Data to Predict Stress in Students

Johannes Simon Fischer

Supervisor: Dr. Simone Tassani

July 2021

Master Thesis on Intelligent Interactive Systems

Universitat Pompeu Fabra

Selecting Features of Breathing & Stability Data to Predict Stress in Students

Johannes Simon Fischer

Supervisor: Dr. Simone Tassani

July 2021

upf. Universitat Pompeu Fabra *Barcelona*

Content

Acknowledgments

A special thanks go to my family and friends who have been there for me unreservedly throughout my university years and have supported me at all times. Furthermore, I would like to acknowledge and express my gratitude to Paula, Klara, and Clara, who made this work possible in this short time by acquiring the data. Finally, I would like to thank my supervisor Dr. Simone Tassani for his guidance and patience throughout this whole project, and that he gave me the opportunity to be part of the BYMBOS project.

Abstract

Repeated or long-time exposure to psychological stress is a risk factor for mental and physical health. Especially students are exposed to stress and lacking stress management skills make them particularly vulnerable to stress-related illnesses. The *Breathing Dynamic Modelling for Body Mind Interaction in Students* (BYMBOS) project tries to tackle this problem by investigating the triangular relationship between breathing, posture, and psychological stress. This exploratory study contributes to this project by selecting discriminative physiological breathing and stability features that can detect stress. For this purpose, four feature selection methods were applied, namely a Genetic Algorithm, a Decision Tree, a Correlation-based, and a Relief-based feature selection approach. For the classification a Support Vector Machine (SVM), as well as the k-nearest Neighbor (KNN) classifier were used. Stability and respiratory data were recorded before and after relaxation was induced through deep breathing and muscle relaxation exercises. The application of the feature selection methods on the physiological data confirmed that including feature selection as a preprocessing step can not only reduce the number of variables significantly but can also increase accuracy, due to the elimination of noise. Predicting stress based on breathing and stability data achieved a *leave-one-subject-out* (LOSO) mean accuracy of up to 85.42%. Stability measures were found to boost the predictive power but were not sufficient for an effective prediction. However, these results may be biased as the relaxation inducing intervention included deep breathing exercises and several subjects were not able to perform deep breathing correctly. Hence, a different labelling procedure based on whether a person was able to perform deep breathing or not, was proposed. The prediction with this labelling solely based on stability measures achieved a LOSO mean accuracy of up to 87.76%. This suggests a relationship between stability, breathing, and psychological stress. Nonetheless, further research with more samples, and procedures to validate the perceived stress in subjects is required to support this hypothesis.

Keywords: Stress Inference; Stability; Breathing; Machine Learning; Feature Selection; Physiological Measures; Stabilometry; Stress Classification.

1. Introduction

Psychological stress is often described as an unspecific response to situations or stimuli (stressor) which the individual perceives as uncontrollable, unpredictable, challenging, or anticipates negative physiological or psychological consequences from it [1]–[3]. The individuals reaction to a stressor is most often perceived as a negative experience, although the psychological and physiological responses aim at helping the individual coping with the stressors and hence, are often beneficial [3], [4]. Through activation of the Hypothalamic-pituitary-adrenal (HPA) axis, circulatory cortisol levels increase, which suppresses immune system activity temporarily to preserve energy [5]. Combined with increased sympathetic nervous system activity [6] this enables the individual to cope with special situations [3]–[5].

However, long time or excessive exposure to stress is related with increasing risk for mental and physical health [3]–[5], [7], [8]. For example, in the psychological domain chronic stress is considered as risk factor for burnout, major depression, and anxiety disorders [3], [8], [9] and from the physiological perspective it is associated, either as risk factor or elevator, with hypertension and cardiovascular diseases [10], and may also increase the risk of atherosclerosis, cancer, stroke, and inflammation [8], [10].

But also muscular and posture related problems are associated with stress. Hauke et al. (2011) conducted a meta-analysis of longitudinal studies and concluded that among other factors psychological distress plays a causal role in the development of musculoskeletal disorders, such as muscular pain in shoulder, neck, or lower limbs. Prins et al. (2008) found that stress is one of the six most common factors among children and adolescents associated with the development of musculoskeletal pain. According to them, the prevalence of musculoskeletal pain in childhood ranges from 21.3% to 34%, whereas the prevalence in adolescence is even higher, ranging from 44% to 47% [12]. Individuals in this age are often school learners or students, using computers which further increases the risk of musculoskeletal pain due to bad sitting posture [13], [14]. Additionally, school and university students are prone to psychological stress. Regehr et al. (2013) conducted a meta-analysis on stress management interventions among students and found that approximately half of the students report stress-related health concerns, and Pascoe et al. (2020) report that according to the Organization for Economic Co-operation and

Development (OECD) 66% of the students reported feeling stressed. These findings indicate a close relationship between posture and psychological stress, which is especially true for students, as they are particularly vulnerable to incorrect sitting habits and stress.

Of particular concern are therefore methods to tackle the individual stress response and thus, long-term harmful effects. A recent meta-analysis found behavioral and mindfulness-based interventions to be notably efficient to reduce stress in students [15]. The authors concluded that this does not only apply for self-reported stress levels, but also for biological markers, such as salivary cortisol levels. Similar findings were reported in the meta-analysis by Zou et al. (2018) who found that mind-body exercises significantly modulate the sympathetic-vagal balance in adults and also significantly reduce perceived stress. Besides muscle relaxation, this kind of interventions often implement breathing practices that help the individual to focus and control his/her breathing patterns [15], [18], [19]. Diaphragmatic breathing, frequently also called *deep breathing*, involves the deepening of inhalation while expanding the abdomen, and thereafter exhaling slowly while contracting the abdomen [19], [20]. Number of breathing cycles per minute are controlled during deep breathing and vary between four to six [19], [20]. As a comparison, subjects in a resting control group of the study of Ma et al. (2017) had an average respiratory rate of 17 breaths per minute. The advantages are that these breathing practices can be easily taught, are non-pharmacologic, self-administered, and come at no cost. Furthermore, there is growing evidence for the beneficial effects on both, physical and psychological health parameters. Studies show that deep breathing is able to improve the mood [7], increase sustained attention and cognitive performance [19], and reduces the subjects perceived stress [7], [19]. It can moreover decrease the heart rate [7] and has a positive effect on the salivary cortisol level [7], [19]. These findings suggest a tight link between breathing and stress.

In summary the literature indicates that psychological stress is associated with both, posture and breathing. However, only few is known about the triangular relationship between stress, breathing, and posture in general. The *Breathing Dynamic Modelling for Body Mind Interaction in Students* (BYMBOS) project targets this problem with a multi-disciplinary approach, by following two main goals. Firstly, it tries to provide evidence of the triangular link between breathing, stress, and posture, and secondly, attempts to dynamically model the relations which then allows constant monitoring, prediction, and control for future situations [21]. The project especially focuses on students, as this population is particularly exposed to stress with lacking preventative, self-regulatory measures [15], and is also prone to improper postural habits, which lay the ground for musculoskeletal disorders [12]. Additionally, students may benefit the most from increased sustained attention [19], as this could enhance their academic performance. This thesis contributes to the BYMBOS project with the creation of a stress detection model, based on breathing and stability data. For successful modelling, two things need to be addressed. Firstly, the most predictive variables among a variety of breathing and stabilometry measures need to be determined, and secondly, a machine learning model based on the selected variables needs to be trained.

Reducing the number of variables is a well-known problem, especially in studies that aim at predicting stress from multimodal physiological data [4], [22]–[31]. The reduction of input variables, if well performed, is associated with several benefits. First of all, less variables mean less computational cost and thus, enhance learning efficiency [32], [33]. This is particularly important as some variables may just be noise, that can even reduce the predictive performance of a model and lead to overfitting [34], [35]. The elimination of redundant and irrelevant features can also help to increase the comprehensibility, as relations of fewer variables are easier to interpret [34], [35], which is particularly important for this work. A common approach to reduce the number of variables is the application of feature extraction methods which create new features by transforming the original signals or statistical descriptors [33], [34]. Even though these methods are well suited to reduce the number of input variables, the resulting features are often lacking interpretability. For example, the *Principal Component Analysis* (PCA) is known to significantly reduce input dimensions, but the resulting factors of variation can hardly be interpreted. As the goal of BYMBOS is the investigation of the relationship between breathing, posture, and stress, interpretation of the underlying features is crucial and thus, a PCA not applicable. More appropriate approaches are feature selection techniques, whose goal is to identify and select the most informative features and discard irrelevant ones [32]–[36]. One key advantage of feature selection methods, in contrast to feature extraction methods, is the maintained interpretability of data [33], [34].

Feature selection techniques can be classified in three different types: Filter methods, Wrapper methods, and embedded methods. Filter methods use data-intrinsic properties to select features, and thus, are independent of the classification algorithm [32]–[36]. In this thesis two filter methods are applied. Firstly, a *correlation-based feature selection* method, which evaluates a feature subset based on data intrinsic correlations [36]. And secondly, a *relief-based feature selection* method, which individually evaluates and ranks features [33]. The second type of feature selection techniques are wrapper methods. These methods evaluate different feature subsets based on their performance in the modelling algorithm [33], [34], [37]. As an exhaustive brute-force approach is computationally costly, the *Genetic Algorithm* provides an evolutionary-based heuristic, which guides the search. The last type of techniques are embedded methods, which incorporate the feature selection in the inductive algorithm itself [32]–[34]. The most famous example are Decision Tree based algorithms [33], which also come with the advantage of being easily interpretable, and are therefore well suited for this work. To sum up, four algorithms are applied: Correlation-based feature selection, relief-based feature selection, genetic algorithm, and a Decision Tree classifier. All algorithms will be explained in detail in the methods section of this work.

Finally, there will be the predictive modelling step, which includes the training of a machine learning model. The task of this thesis is a binary classification problem, which tries to predict if a person is stressed or not based on breathing and stabilometry data. Hence, supervised machine learning algorithms can be applied. A literature review on related work (further described in section 2) revealed that the most common algorithms to detect stress based on multimodal physiological data are Support Vector Machines (SVM), Decision Trees, and the k-nearest neighbor method [4], [5], [22], [26]–[29], [38], [39]. Therefore, these three machine learning algorithms are applied to detect stress in students. Further justification can be found in section 3.4.

The main objective of this thesis is therefore, to identify the most relevant breathing and stabilometry features, that are able to predict stress. In the following section related work about stress detection based on physiological data is presented. Thereafter, the data acquisition procedure is described, and each feature selection and classification algorithm further explained. The results are then presented and lastly, further discussed.

2. Related Work

Several works in recent years applied machine learning models to predict acute stress based on physiological data.

One of the first highly relevant studies aimed at detecting stress was conducted by Healey & Picard (2005). The participants of this study were following a car-driving protocol with three different situations, each inducing a different stress-level. The low stress condition consisted of resting in the car without driving and was used as baseline. High stress was induced letting the subjects drive on a busy main street in the city with stop-and-go traffic and unexpected hazards, like cyclists and careless crossing pedestrians. The third part of the route was driving on a highway and ought to induce a medium level of stress. A questionnaire analysis confirmed that these different induced stress levels were perceived as intended. During the experiment physiological signals were monitored using five sensors: electrocardiogram (ECG), electromyogram (EMG), galvanic skin response (GSR) on hand and feet, and respiration. From the obtained signals a 22-dimensional feature vector was computed and projected onto a two-dimensional space using Fisher's projection matrix. Classification was performed using a linear discriminant analysis, resulting in an accuracy of 97.4%. As the projection of the feature vector onto a smaller space eliminated interpretability, the authors conducted a further correlational analysis, that identified skin conductivity and heart rate measures as the best correlates of stress.

The dataset acquired by Healey & Picard (2005) was subject of several other scientific analyses in recent years (e.g., [23], [27]). For example, Ghaderi et al. (2016) performed a literature-driven feature extraction resulting in a total of 78 features. Feature selection was performed using Weka software [40] and the built-in feature ranking method *InfoGainAttributeEval*[1]. Classification was then conducted using a SVM and a k-nearest neighbor algorithm, based on several selected feature subsets of different length. The SVM yielded the highest accuracies, with 98% for a subset containing 20 features, and 96% for a subset with 8 features, respectively. The accuracy of using all 78 extracted features ranged from 85% to 93% and were all below the accuracy of the best subsets. These findings further support the effectiveness of feature selection methods as

[1] Documentation weka: https://weka.sourceforge.io/doc.dev/weka/attributeSelection/InfoGainAttributeEval.html.

preprocessing step for machine learning models. Furthermore, the authors concluded that the respiration sensor is the most important sensor for stress detection.

Napoletano & Rossi (2018) used a dataset acquired using a similar paradigm as the one introduced by Healey & Picard (2005). In contrast to the original paradigm, the driving happened in a simulator and more subjects participated (52 vs. 24). For the analysis Napoletano & Rossi (2018) extracted different types of features from heart rate (HR) and breathing rate (BR) measures and evaluated the predictability using a SVM. HR and BR features were evaluated separately, as well as the combination of both. The authors concluded that the breathing rate is more robust in detecting stress than the heart rate, but that the combination of both outperformed all (63%, 61%, and 69% accuracy, respectively).

A different stress-inducing paradigm was used by Shi et al. (2010). Besides rest periods, the 22 subjects were exposed to three different types of stressors: public speaking, mental arithmetic tasks, and a cold pressor stressor, which aimed at representing social, mental, and physical stress, respectively. The stress annotation happened by collecting Ecological Momentary Assessment (EMA) results before and after each stressor and resting period. This assured that the applied stressors worked as intended. During the lab study participants wore a chest band containing the following sensors/measures: ECG, GSR, respiration, and temperature. The authors then used a SVM and incorporated person-specific information into the model which resulted in 62% precision at 80% recall.

Plarre et al. (2011) induced stress similarly as Shi et al. (2010) to 21 subjects, but only used heart rate measures and respiration as input variables to their stress detection model. Validation of stress induction was done by using self-reports. The authors used three classifiers: SVM, Decision Tree, and AdaBoost. To select the most distinguishing features, the authors applied a correlation-based feature selection algorithm, as it is also applied in this work. The resulting subset consisted of 13 features in contrast to 35 features, and increased accuracy of the model, but according to the authors, not significantly. Furthermore, Plarre et al. (2011) evaluated the impact of feature normalization on accuracy and concluded that normalization increases accuracy (e.g., Decision Tree 82.3% vs. 87.6%, respectively). The highest accuracy was achieved using

AdaBoost (90.1%). The authors concluded that respiratory features are highly discriminative of stress.

Martin & Diverrez (2016) acquired a dataset of 24 participants by inducing stress through time pressure on arithmetic tasks. Each participant underwent two conditions, one declared as stressful, and the other declared as not stressful. The stressful situation included an arithmetic task which had to be answered before a signal occurred 900 milliseconds after the presentation of the calculation. For the situation declared as not stressful, a calculation was presented as well, but participants had to answer following a beep occurring 2650 milliseconds after calculation onset. To validate the effect of these two conditions on the perceived stress the authors used two standardized scales: The Short Stress State Questionnaire (SSSQ), which evaluates the three stress aspects Engagement, Distress, and Worry, as well as the Raw-TLX (RTLX), which measures the perceived workload on the six dimensions mental demand, physical demand, temporal demand, performance, effort, and frustration. All dimensions of the RTLX and the Distress aspect differed significantly between the two conditions, which is why the authors concluded that stress induction has been successful. During both tasks electrodermal activities, as well as heart rate, and respiration were measured and used as input features for classification. The authors tried four different classification approaches: Random forest, Naïve Bayes, Neural Network, and SVM. The Random forest classifier achieved the highest accuracy with 73%.

Riera et al. (2012) used Fishers Discriminant Analysis (FDA) to predict stress based on electroencephalography (EEG) and EMG measures and achieved an accuracy of 79%. Stress was induced using the Trier Social Stress Test [43], the Stroop Color Word Task, as well as arithmetic tasks, reading, and an unexpected faked blood sample extraction. No self-reports were used to validate the success of stress induction, but the authors mapped EEG information into the valence and arousal model of emotions, which indicated that all tasks tended to high arousal and negative affect. However, due to a small sample size of 12 participants, interpretations should be made with caution.

Bobade & Vani (2020) used the publicly available WESAD dataset [44] which provides physiological data from 15 participants in three affective states: neutral, stress, and amusement. Stress was induced using the Trier Social Stress Test [43] and has been validated using four questionnaires: the SSSQ, the Positive and Negative Affect Schedule (PANAS), the State-Trait Anxiety Inventory (STAI), and the Self-Assessment Manikins (SAM). Based on heart rate measures, respiration, GSR, temperature, and three-axis acceleration Bobade & Vani (2020) trained seven different classification models, which were: SVM, k-nearest neighbor, Linear Discriminant Analysis, Random Forest, Decision Tree, AdaBoost, and Deep Neural Network. The neural network achieved the overall best performance with an accuracy of 95%, directly followed by the SVM with an accuracy of 93%. To reduce the amount of input variables the authors applied a *Principal Component Analysis* and took the 20 first principal components. Again, due to the small sample size interpretation should be made with caution, as the model may be prone to overfitting with such a small sample size.

Abouelenien et al. (2016) overcame this problem by recording physiological measures from 50 subjects. Stress has been induced by different speaking scenarios and self-assessments of the subjects, whether a task was most or least stressful, were used as ground-truth labelling. The authors measured GSR, blood volume pulse, abdominal respiration, as well as several thermal features and modeled the classification problem using a Decision Tree and fused measurements in multiple different combinations. The fusion of thermal and respiratory features outperformed all other single modality subsets, as well as all other combinations of modalities, and achieved an accuracy of 75%.

This literature review of papers detecting stress revealed that Support Vector Machines (SVM), Decision Trees, and k-nearest neighbor (KNN) method were the most used classification algorithms. For example, SVMs were used in [4], [5], [22], [27], [28], [38], [39], [41] and achieved accuracies on stress detection from 61% [41] up to 98% [27]. All the reviewed papers using SVMs had respiration and heart rate measures as input variables [4], [22], [27], [28], [39], [41]. Further used sensor signals were skin conductivity [4], [22], [27], [39], temperature [4], [22], [28], [39], and EMG signals [27], [28].

Decision Trees, or derivations of it, were used in [4], [28], [29], [39]. Derivatives were, for example, a Random forest classifier, which is an ensemble of Decision Trees, and has been applied by [39] as well as AdaBoost, which consists of a cascade of several weak Decision Tree classifiers, with each classifier often just using one feature to split the data, used by [28]. Accuracies on stress detection with Decision Trees were in the range of 73% [39] to 90% [4]. All studies using a Decision Tree classifier used heart rate measures, respiration, and skin conductivity as features [4], [28], [29], [39]. Moreover, temperature [28], [29] and three-axis acceleration were used [4], [28].

Another classification algorithm that has often been used is the k-nearest neighbor (KNN) method [27], [28], [39], which also achieved remarkable results (e.g., 95% [27]). Several other algorithms were also used in detecting stress based on physiological measures. Two studies trained a neural network to predict stress [28], [39], and achieved accuracies up to 95% [28]. Densely connected neural networks have several parameters to train, which get updated through feeding forward data and optimizing the parameters using backpropagation [51]. One common problem is overfitting, where the parameters are tuned to fit to the training data, but are not capable of predicting new, unseen samples. Big sample sizes are usually used to overcome this problem, but as sample sizes in experiments with physiological measurements are often small, densely connected networks may not be applicable. Another algorithm often used, was random forest which tends to increase predictability, but at cost for interpretability. As the focus in this thesis is on interpretability, and Decision Tree Classifiers are easier to interpret, this is the algorithm of choice. Lastly, Linear Discriminant Analysis and Fishers Discriminant Analysis should be mentioned, as they yielded accuracies up to 90% [28] and 97% [24], respectively. Even though the results are promising these algorithms are using a linear combination of features, which may not be applicable in this thesis work due to non-linear relationships. This reason also applies for Logistic Regression [51], which has also not been described in any reviewed paper as potential stress detection algorithm.

Taken together, the above-described findings suggest that stress detection applying machine learning techniques such as SVM, KNN, and Decision Trees is possible and that physiological measures are well suited as features. This is further confirmed by a recent survey on applied machine learning methods on physiology based stress detection systems [26]. According to the described studies, especially respiration seems to be a promising and commonly used indicator of stress [4], [27], [29], [41]. However, none of the studies included stabilometry data as predictors for stress. Three studies used EMG signals as predictor [24], [28], [42], but no conclusions were drawn about postural stability. Therefore, there are two main contributions of this work. Firstly, it contributes by including stabilometry measures as input features for stress detection. And secondly, it focusses not only on prediction, but also tries to identify the most discriminative features that can predict stress.

3. Methods

3.1 Subjects

For this thesis, the dataset consists of physiological recordings from 25 healthy volunteering students. 18 participants were male, and the mean age was 23.2 years (standard deviation: 3.34). Data from ten participants, all male, was obtained from a previous study conducted in the same motion capture laboratory, following the same protocol. The mean age of this sample was 28 years (standard deviation: 7). The remaining data samples ($n = 15$) were recorded as part of the BYMBOS research project, with 7 female participants and an average age of 20.0 years (standard deviation: 1.75). All participants gave informed written consent prior acquisition. Data from one participant was incomplete.

3.2 Measures

Stabilometry. The postural control system is associated with keeping the balance during quiet standing [18], [45]. Swaying during quiet standing indicates reduced stability of a person [18], which can be measured with a force plate. For this study a BTS P-6000 force plate (500 Hz sampling, BTS S.p.A., Milan, Italy) was used to measure the foot-ground forces. The most common characterization of postural steadiness is the anterior-posterior (AP) and medial-lateral (ML) displacement of the center-of-pressure (COP), which is the point of application of the vertical reaction vector on the force plate surface [45]. This is further illustrated in Figure 1. From the AP_O and ML_O time series several features according to [45] and [18] were extracted. A summary and short description of the measurements can be found in Table 1. Computations were made following the methods of [45].

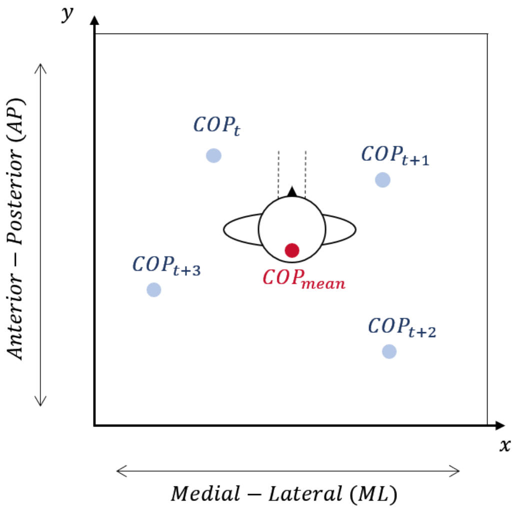

Figure 1. Postural steadiness measured with anterior-posterior (AP) and medial-lateral (ML) displacement of the center-of-pressure (COP). Blue dots represent the COP at different time steps, the red dot is the mean COP.

Table 1. Brief description of the stabilometry features extracted from the force plate signal. Units are mm (millimeters), s (seconds), cm (centimeters), Hz (Hertz), deg (degree), AD (adimensional).

Measure	Unit	Description
Time-domain stability measures of sway		
Transversal COP displacement	mm	COP displacement along X axis
Longitudinal COP displacement	mm	COP displacement along Y axis
Radius	mm	Mean COP distance from the barycenter
Transversal range		Difference between maximum and minimum Px values
Longitudinal range		Difference between maximum and minimum Py values
Trace length	mm	Sum of the COP trace on the platform
Speed	mm/s	Mean of the COP displacement velocity
LFS	mm^{-1}	Length as a function of surface
Equivalent area	mm^2	Area marked out by the COP
Equivalent radius	mm	Radius of a circle with an area equal to the equivalent area
Inertial axes	mm	X and Y axes around which the subject is swaying
Regression angle	deg	Inclination of the X inertial axis in regard to the x axis of the reference system
Geographic area	mm^2	Inertial ellipse area
Sway density radius	mm	Radius of the time instants of the sway density curve in which the subject had peaks of stability
Peak number		Number of peaks of the sway density curve
Peak amplitude	s	Mean and SD of the peaks of the sway density curve
Peak time	s	Mean and SD of the time between peaks of the sway density curve
Peak distance	mm	Mean and SD of the mean distance between peaks of the sway density curve
Frequency-domain stability measures of sway		
Maximum Peak (Px, Py)	AD	Maximum peak of the COP frequency spectrum in the x and y direction, respectively
Maximum Frequency (Px, Py)	Hz	Frequency of the maximum peak of the spectrum in the x and y direction, respectively
Mean Amplitude (Px, Py)	AD	Mean amplitude of the spectrum in the x and y direction, respectively
Mean Frequency (Px, Py)	Hz	Mean frequency of the spectrum in the x and y direction, respectively
Maximum Peak D	AD	Maximum peak of the spectrum of the distance from the barycenter
Maximum Frequency D	Hz	Frequency of the maximum peak of the spectrum of the distance from the barycenter
Mean Amplitude D	AD	Mean amplitude of the spectrum of the distance from the barycenter
Mean Frequency D	Hz	Mean frequency of the spectrum of the distance from the barycenter

Breathing. Optoelectronic plethysmography (OEP) has been proven to be an accurate, non-invasive, and robust measurement of chest wall kinematics during respiration [46]–[48]. From this movements, volume variations of the chest wall and its different thoraco-abdominal compartments can be obtained to provide the complete ventilatory pattern. The three compartments used for this thesis are the pulmonary rib cage (RCp), abdominal rib cage (RCa), and the abdomen (AB). Prior to tracking several reflective markers were placed on different anatomical parts of the subject according to the protocol of [48]. 89 markers were placed over specific parts of the chest (illustrated in Figure 2). During recording eight infrared light video cameras (Smart-DX 700, 1.5 Mega Pixels, 250 frames per second, BTS S.p.A., Milan, Italy) tracked the position of the markers using the SMART Capture software developed by BTS Engineering (BTS S.p.A., Milan, Italy). Thereafter, the 3D coordinates of the markers were computed by integrating the information of the different cameras to obtain a 3D model and then, each marker was labelled manually. Finally, the rate of the tidal volume, which is the amount of air moving in or out of the lungs during a normal breath, for each of the three compartments, as well as the total tidal volume were extracted and the mean and standard deviation of each of them computed. Hence, eight features regarding breathing were obtained from the OEP signals. Additionally, a new feature was created by subtracting the pulmonal ribcage volume from the abdominal volume. This feature indicates whether a person was breathing abdominally or not.

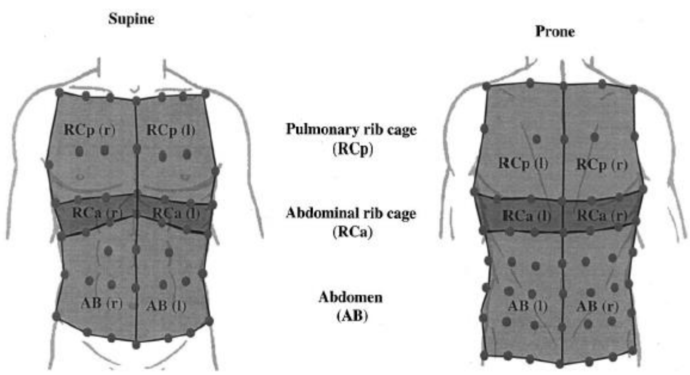

Figure 2. Positioning of the markers on the three different chest wall compartments (Pulmonary rib cage, abdominal rib cage, and abdomen. Left: Supine, Right: Prone. Adapted from [48].

Skin conductivity. The electrodermal response (EDA), or often also called galvanic skin response (GSR), is a psychophysiological indicator of a person's emotional arousal [49], [50]. If a person experiences an arousing stimulus, the electrical properties of the skin change, due to the production of sweat. The sweat secretion is directly proportional to the level of arousal, with more arousing stimuli leading to higher sweat secretion [49]. The most common measurement of this is the skin conductance (SC) [50]. For that two electrodes are placed on the skin of a subject and an electrical potential applied. As sweat is an efficient conductor of current, SC increases with 1 to 4 seconds latency, as soon as the subject perceives a stimulus as personally significant [49]. In this study the SC was measured on the fingers of the subjects with a Shimmer3 GSR+ unit (Shimmer, Dublin, Ireland), as there is a high concentration of sweat glands [50]. This has been done to indicate whether a subject experiences stress or not. For every subject the mean and standard deviation were extracted from the signal, with a lower signal cut-off limit of 0.5 micro siemens (µS). However, due to delays in data acquisition this data was only available for 17 subjects.

3.3 Procedure

All data has been obtained at the Motion Capture Lab at Pompeu Fabra University, Barcelona. Before the start, participants were informed about the general structure of the acquisition and that they can leave the experiment without any justification at any time, which they had to confirm through a signed informed consent. Thereafter participants were asked to change, so that men only wear short pants, and women short pants and any kind of brasserie (e.g., sport top or upper part of a swimming suit). After measuring height and weight, the lab assistant then performed palpation to identify bones and anatomical landmarks to correctly place the reflective markers for the optoelectronic plethysmography. Electrodes for the measurement of skin conductivity were placed on the fingers of the participants and the sensor device attached to the wrist.

The experiment consisted of four different conditions as proposed by Tassani et al. (2019). Standard quiet standing with open and closed eyes, and relaxed quiet standing with open and closed eyes. In each condition subjects were instructed to stand quiet and with the feet about as far apart as the hips on the force plate in the middle of the room. In the open eyes condition, they were further asked to look at a target on the wall approximately at the height of their eyes. Each recording lasted approximately 180 seconds and stabilometry, breathing, as well as skin conductivity were recorded. Firstly, the standard quiet standing, open and closed eyes, was performed. Thereafter the lab assistant guided the participants through some muscular relaxation and deep breathing exercises. Subsequently, the previous recording was repeated, and subjects were asked to stand quiet on the force plate, one time with open and the other with closed eyes. The procedure is illustrated in Figure 3. To verify that the relaxation and deep breathing exercises induced relaxation in the participants a two-way repeated measures ANOVA has been conducted, with the galvanic skin response average being the dependent variable, and eyes (open vs. closed), as well as state (normal vs. relaxed) being the two within-subject factors.

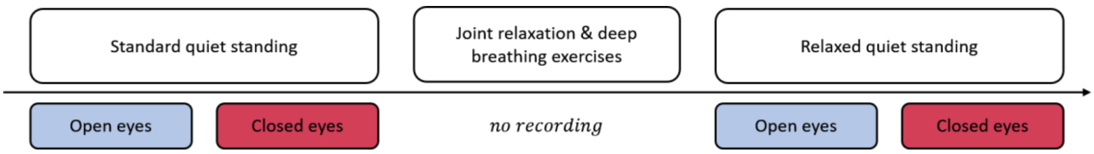

Figure 3. Procedure of the experiment. Firstly, standard quiet standing with open and closed eyes was performed and breathing, stabilometry, and skin conductivity recorded. Secondly, the subject was guided through muscle relaxation and deep breathing exercises, and thirdly, the measurement of the beginning was repeated.

The feature set for classification contained all extracted breathing and stabilometry variables. The classification task of this thesis consisted of predicting the condition standard quiet standing versus relaxed quiet standing, which is labelled as stressed or not stressed, respectively. The assumption that participants were (more) stressed in the first condition is justified by the fact that participants were in a completely new environment, with new social interactions, and had to stand quiet half naked in front of people they didn't know. Thereafter, relaxation was induced by the muscle relaxation and deep breathing exercises, hence, it is assumed that in the second condition, relaxed quiet standing, participants were not or at least less stressed.

However, this relaxation induction and labelling procedure comes with several limitations, which are further discussed in section 5 of this thesis. To overcome some of these limitations another labelling procedure has been applied in an exploratory analysis. For that, a subset with only the relaxed state was chosen and samples classified according to the difference between mean abdominal volume (AV) and mean pulmonal ribcage volume (PRV)

$$x = AV - PRV \; . \qquad (1)$$

Samples with values greater or equal to 0 are labelled as not stressed, due to higher abdominal volume, and values less than 0 are labelled as stressed, due to higher volume in the pulmonal ribcage. This indicates whether a person was able to perform deep breathing or not. Input features for this analysis were just stabilometry measures. Since the resulting dataset was unbalanced anyway, the subject with partially missing data was added.

3.4 Classification

The problem of detecting stress vs. non-stress is a binary classification problem. Input variables are different features extracted from breathing and stabilometry measurements, and the target variable is binary, with labels $0 \coloneqq$ *no-stress* and $1 \coloneqq$ *stress*.

Decision Tree classifiers, SVMs, and KNN have been proven to be a good choice when it comes to classification based on physiological measures, such as respiration, heart rate, and skin conductivity, achieving results up to 90% [4], 98%, and 95% [27], respectively. As these are all non-linear methods and they are well suited for the stress detection task pursued in this thesis. Furthermore, SVMs are effective in high-dimensional spaces and are even applicable in cases with small sample size as the optimization problem is convex [51]. Decision Tree classifiers come with the benefits that they are easy to interpret and already embed feature selection. Hence, these three algorithms were used for binary classification in this work. However, the evaluation of feature subsets in the Genetic Algorithm are solely based on SVMs as KNNs come with high computational cost for large feature vectors.

Due to a limited amount of data and the thereby resulting lack of sufficient validation data several studies applied k-fold cross-validation [4], [22], [27]–[29], [39], [41], [42]. This process splits the data into k smaller sets and trains the model in each split on $k - 1$ folds, testing it on the remaining fold. The final accuracy is then the average accuracy over all splits. Figure 4 further illustrates the procedure with four splits ($k = 4$). A special cross-validation method is called *leave-one-subject-out* where the number of folds k is equal to the number of samples n. Hereby the model gets trained with $n - 1$ samples and tested on one single remaining sample. This process is repeated n-times so that each subject has been the subject under test once.

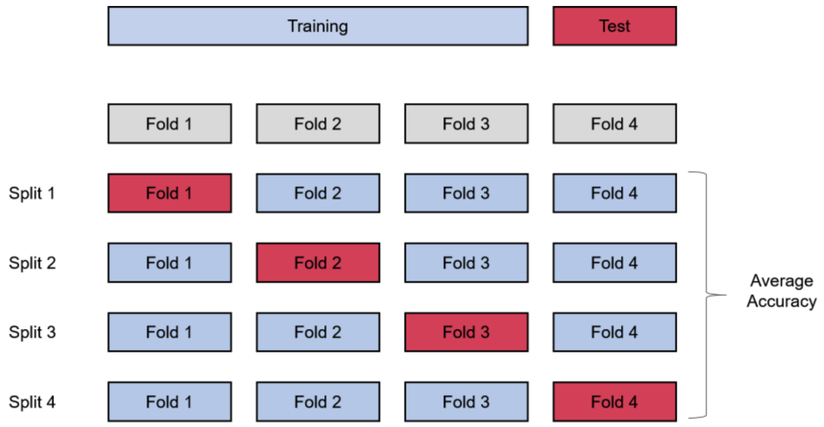

Figure 4. Procedure of k-fold cross-validation with $k = 4$.

As the *Leave-one-subject-out* cross-validation was applied in most of the reviewed studies [22], [28], [29], [41], [42], it is also applied in this thesis. However, for the Genetic Algorithm, where in each epoch a SVM is trained for each of the offsprings of the population, it should be mentioned that due to the high computational cost the *leave-one-subject-out* cross-validation would cause, only 10-fold cross-validation is applied. In the following all three classification algorithms are explained in detail.

3.4.1 Support Vector Machine

Support Vector Machines are a supervised, geometry-based machine learning algorithm that treat datapoints as spatial coordinates [51]. The general goal can be specified as finding a hyperplane in a multi-dimensional feature space that can best separate two classes. A simple two-dimensional example for a linearly separable toy dataset can be found in Figure 5. Let $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n\}$ with $\boldsymbol{x}_i \in \mathbb{R}^d$ be a set of training samples and

$Y = \{y_1, y_2, \dots, y_n\}$ the corresponding labels with $y_i \in \{-1,1\}$. A hyperplane can be specified as

$$\sum_{i=1}^{d} \boldsymbol{w}_i \boldsymbol{x}_i + b = \boldsymbol{w}^T \boldsymbol{x} + b = 0 \tag{2}$$

with $\boldsymbol{w} \in \mathbb{R}^{\mathrm{d}}$ being the weight vector, $b \in \mathbb{R}$ the bias, and $\boldsymbol{x} \in \mathbb{R}^{\mathrm{d}}$ points that lie on the hyperplane. A SVM tries to find $\boldsymbol{w}$ and b so that

$$\begin{aligned} \boldsymbol{w}^T \boldsymbol{x}_i + b > \mathbf{0} \quad & \text{for all } \boldsymbol{x}_i \text{ whose } y_i = +1 \\ \boldsymbol{w}^T \boldsymbol{x}_i + b < \mathbf{0} \quad & \text{for all } \boldsymbol{x}_i \text{ whose } y_i = -1 \end{aligned} \tag{3}$$

what can be rewritten as

$$y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) > \mathbf{0} \tag{4}$$

for all training sample points [51]. In Figure 5 this hyperplane is represented as the dashed grey line in the middle. However, for the parameters $\boldsymbol{w}$ and b there may exist several solutions specifying a hyperplane that can separate both classes. The grey dashed line in Figure 5 can have several different slopes and biases and still may be able to separate the blue from the red datapoints. Hence, it is not only desirable to find a hyperplane that separates both classes, but also to find the one with the smallest generalization error and thus, the maximum distance to any of the datapoints. This is achieved in SVMs by using the concept of *margin*, which is defined as the minimum distance between the decision boundary and any of the training samples. In Figure 5 the margin is illustrated as the black arrow between the dashed line and the line in the middle. For better generalization SVMs chose the decision boundary with maximum margin. The perpendicular distance of a datapoint $\boldsymbol{x}$ from a hyperplane is defined as

$$\frac{|\boldsymbol{w}^T \boldsymbol{x}_i + b|}{||\boldsymbol{w}||}. \tag{5}$$

The only valid solutions are the ones where all datapoints are correctly classified, so that equation (4) holds. Hence, the distance of a point $\boldsymbol{x}_i$ to the decision hyperplane is given by

$$\frac{y_i(\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x}_{\boldsymbol{i}} + b)}{\|\boldsymbol{w}\|}. \tag{6}$$

The goal is to optimize the parameters $\boldsymbol{w}$ and b so that the perpendicular distance of the closest point $\boldsymbol{x}_{\boldsymbol{i}}$ to the decision boundary (margin) is at its maximum. These closest points to the decision boundary are called *support vectors* and are further illustrated in Figure 5 as orbited datapoints. Hence, the solution with maximum margin is found by solving

$$\mathrm{argmax}_{\mathbf{w},\mathrm{b}} \left\{ \frac{1}{\|\boldsymbol{w}\|} \ \mathrm{min}_{\mathrm{n}} \left[y_n(\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x}_{\boldsymbol{n}} + b) \right] \right\} \tag{7}$$

where n should be minimized in order to find the minimum number of support vectors, and the factor $1/\|\boldsymbol{w}\|$ was taken out of the optimization as $\mathbf{w}$ does not dependent on n [51]. For the support vector it applies that

$$y_i(\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x}_{\boldsymbol{i}} + b) = 1 \tag{8}$$

and therefore, all datapoints satisfy the constraint

$$y_i(\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x}_{\boldsymbol{i}} + b) \geq 1 \quad \forall i \ \in \{1, \dots, \mathrm{n}\}\,. \tag{9}$$

Hence, the optimization problem in equation (7) only requires maximizing $1/\|\boldsymbol{w}\|$ which is equivalent to minimizing $\|\boldsymbol{w}\|^{\mathbf{2}}$ [51], [52]. This leaves us with the optimization problem

$$\begin{gathered} \mathrm{argmin}_{\mathbf{w},\mathrm{b}} \frac{1}{2} \|\boldsymbol{w}\|^{\mathbf{2}} \\ \text{s. t.} \quad y_i(\boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x}_{\boldsymbol{i}} + b) \geq 1 \quad \text{for all } 1 \ \leq i \ \leq n\,. \end{gathered} \tag{10}$$

The constant factor ½ is included for mathematical convenience [51]. Equation (10) is a *quadratic programming* problem, which minimizes a quadratic function with multiple linear inequality constraints, and can be further simplified with Lagrange multipliers to a dual optimization problem [51], [52]. The optimization problem is convex, and thus any local parameter solution is also a global minimum [51].

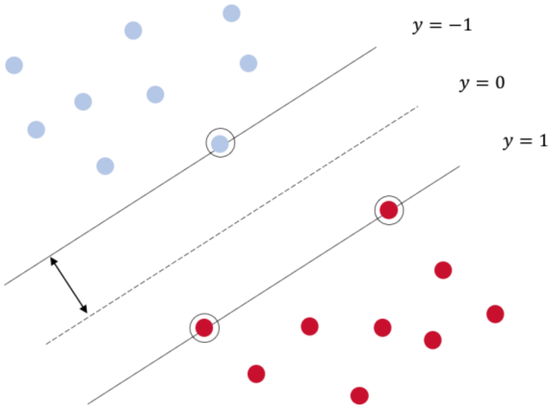

Figure 5. Hyperplane that separates the red and the blue class with maximum margin. The boundary is determined by a subsample of the datapoints, which are called support vectors (indicated by circles around datapoints).

So far, it has been implicitly assumed that data is linearly separable, as illustrated in Figure 5. To cope with data that isn't linearly separable, SVMs follow two strategies.

The first strategy is to introduce so called *slack variables* $\xi_i \geq 0$ with $1 \leq i \leq n$. Therefore, each datapoint has a slack variable with $\xi_i = 0$ for points that are correctly classified (inside or on the correct margin boundary), and $\xi_i = |y_i - \boldsymbol{w}^T\boldsymbol{x_i} + b|$ for other points. For example, all red orbited datapoints in Figure 5 have $\xi_i = 0$, as they are either on the correct margin boundary or inside. A red datapoint between the dashed middle line ($y = 0$) and the red margin boundary ($y = 1$) would have $\xi_i < 1$ and a red datapoint between $y = 0$ and $y = -1$ would have $\xi_i > 1$ [51]. The constraints of equation (9) are therefore replaced with

$$y_i(\boldsymbol{w}^T\boldsymbol{x_i} + b) \geq 1 - \xi_i \quad \forall i \in \{1, \dots, \mathrm{n}\}\,. \tag{11}$$

The optimization problem then changes to

$$C \sum_{i=1}^{n} \xi_i + \frac{1}{2} ||\boldsymbol{w}||^2 \tag{12}$$

where $C > 0$ is a parameter that controls how much incorrect classified samples are penalized and thus, acts as regularization [51], [52]. Hence, high values of C strongly penalize misclassified datapoints, whereas with low C values incorrect classifications are tolerated.

The second strategy to cope with non-linearly separable data is the projection $\phi(\mathbf{x})$ of datapoints into a higher-dimensional space, where data is linearly separable [51]. This is illustrated in Figure 6, where data has been projected from two to three dimensions. In this higher-dimensional space the previously defined linear SVM model could be applied by adapting the constraints from equation (11) to

$$y_i(\boldsymbol{w}^{\boldsymbol{T}}\phi(\boldsymbol{x_i}) + b) \geq 1 - \xi_i \quad \forall i \in \{1, \ldots, \mathrm{n}\}\,. \tag{13}$$

with $\phi: X \rightarrow \mathbb{R}^d$ being the projection. However, usually feature vectors are large and mapping each datapoint $\boldsymbol{x}$ into a higher dimension and then finding the boundary within this higher-dimensional space is computationally costly. Here, the *kernel trick* comes into place and allows to perform computation in the feature space instead of the higher-dimensional space by representing relations between pairwise data samples rather than explicitly applying transformation. The kernel function takes vectors from the original space as inputs and returns the dot product of them in the higher dimensional space. This can be specified as

$$k(\boldsymbol{x_i}, \boldsymbol{x_j}) = \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x_j}) \tag{14}$$

with $\boldsymbol{x_i}, \boldsymbol{x_j} \in \boldsymbol{X}$ [51]. Hence, in the training algorithm only $k(\boldsymbol{x_i}, \boldsymbol{x_j})$ needs to be applied without the need to have explicit knowledge about $\phi(\boldsymbol{x})$ [52].

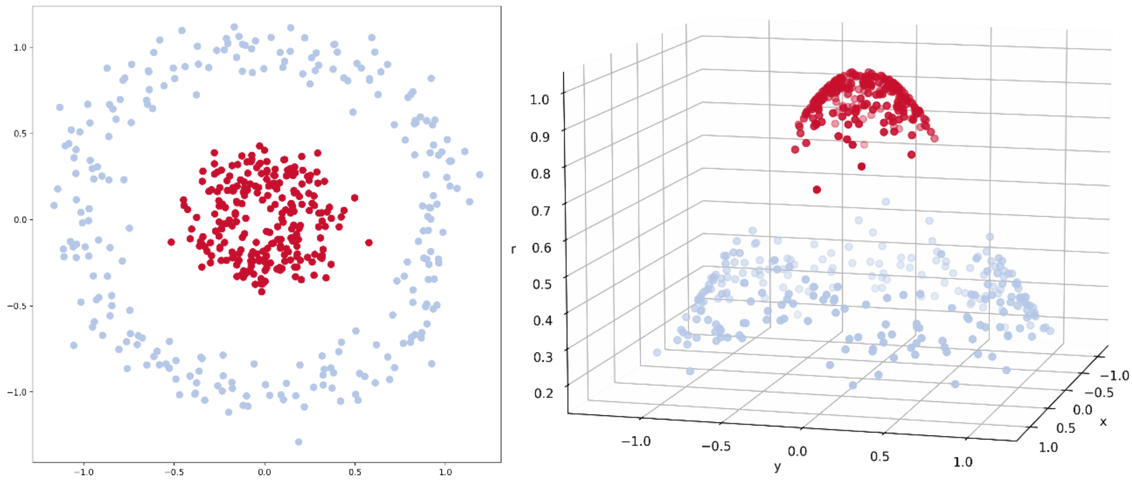

Figure 6. Non-linearly separable data (left) projected into a higher dimension (right) to be linearly separable by a hyperplane

For this thesis the *radial basis function* (rbf) kernel[2]

$$k(\boldsymbol{x_i}, \boldsymbol{x_j}) = exp\left(-\left\|\boldsymbol{x}_i - \boldsymbol{x}_j\right\|^2 / 2\gamma^2\right) \tag{15}$$

has been applied, with $\gamma > 0$ defining how far the influence of a single datapoint reaches, which is another parameter that needs to be tuned.

A grid-search has been used in this thesis to determine the best parameters C and γ. However, as this process is computationally costly it has only been applied in the final analysis and not during the feature selection process with the Genetic Algorithm. For this, C and γ have been set to constant values. The grid-search and classification in this thesis have been performed with Python's sci-kit learn implementation [53].

3.4.2 k-Nearest Neighbor

In contrast to SVMs, where the goal is to find a hyperplane that separates the classes, the k-nearest neighbor method tries to approximate the underlying distribution of the data. Intuitively spoken, classification of a new sample happens by first determining the k-nearest neighbors using a distance measure, and then, classifying it using a voting mechanism, so that the new sample gets the label that is most prominent in its neighborhood [54].

Let again $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n\}$ with $\boldsymbol{x}_i \in \mathbb{R}^d$ be a set of training samples and $Y = \{y_1, y_2, \dots, y_n\}$ the corresponding labels with $y_i \in \{0, 1\}$. Let further $\rho\colon\ X \times X \ \rightarrow \mathbb{R}$ be a function that returns the distance between any two datapoints $\boldsymbol{x_i}, \boldsymbol{x_j} \in X$. It is then possible to order for each $\boldsymbol{x} \in X$ the other samples $\boldsymbol{x}_j \in X/\boldsymbol{x}$ according to their distance $\rho(\boldsymbol{x}, \mathbf{x}_\mathrm{j})$, so that

$$\rho(\boldsymbol{x}, \mathbf{x}_\mathrm{j}) \ \leq \cdots \ \leq \rho(\boldsymbol{x}, \mathbf{x}_\mathrm{n}) \quad \forall j \in \{1, \dots, n-1\}\,. \tag{16}$$

Classification then happens using the parameter k which determines how many neighbors are involved in the decision, so that for a datapoint $\boldsymbol{x} \ \in \boldsymbol{X}$ the label is the majority label

[2] Documentation sci-kit learn:
https://scikit-learn.org/stable/modules/generated/sklearn.gaussian_process.kernels.RBF.html

among $\{\mathbf{y}_j : j \leq k\}$ [54]. Figure 7 illustrates the decision boundary for a binary classification problem with different values for k.

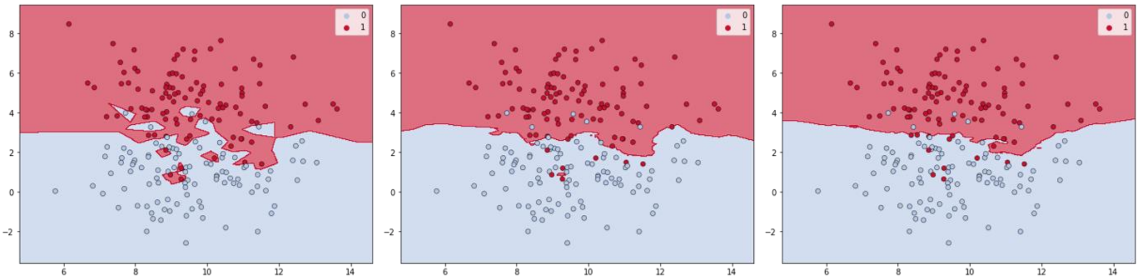

Figure 7. Decision boundary for a binary classification problem with different values for k. From left: $k = 1$, $k = 5$, and $k = 10$.

For this research work the Euclidean distance has been used as distance measure which is defined as

$$\rho(\boldsymbol{x}, \boldsymbol{x}') = \|\boldsymbol{x} - \boldsymbol{x}'\| = \sqrt{\sum_{i=1}^{d}(x_i - x_i')^2} \tag{17}$$

with $\boldsymbol{x}, \boldsymbol{x}' \in X$. Due to the small sample size and following the recommendations of sci-kit learn[3] the brute-force search has been applied for finding the distances, computing all pairwise distances of the dataset. Furthermore, for this work, the neighbors are weighted using the inverse of their distance, so that closer datapoints have a greater influence on classification than neighbors that are further away. The impact of weights is further illustrated in Figure 8. Classification has been performed using Python's sci-kit learn implementation [53].

[3] Documentation sci-kit learn: https://scikit-learn.org/stable/modules/neighbors.html

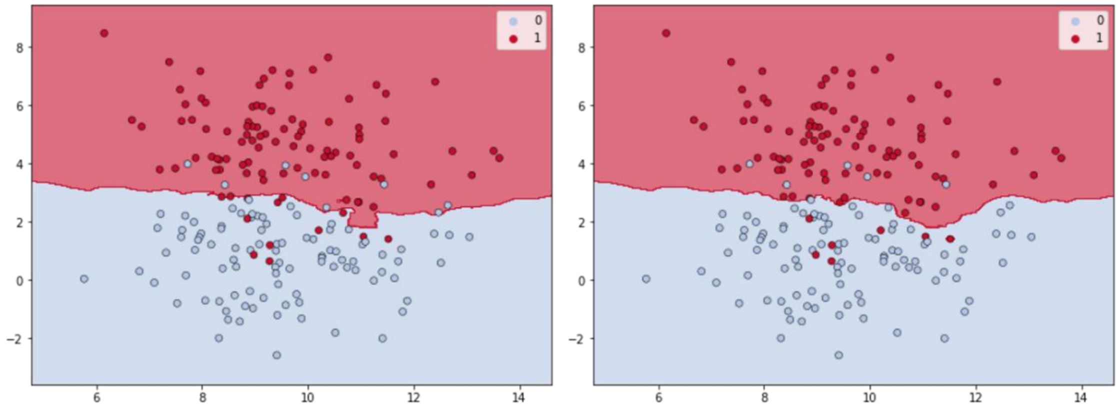

Figure 8. Left: Decision boundary for $k = 15$ without weights. Right: Decision boundary for $k = 15$ with each neighbor being weighted by the inverse of its distance.

3.4.3 Decision Tree

Decision Trees infer decision rules from features of a dataset and create a binary Decision Tree based on these rules. Each node of a tree represents a test on a given feature and each branch is the outcome of this test. The classification happens by going down the tree from the root node until a leaf node is reached, which represents a label or decision. Figure 9 illustrates a Decision Tree for a toy example, to classify whether a person is stressed or not.

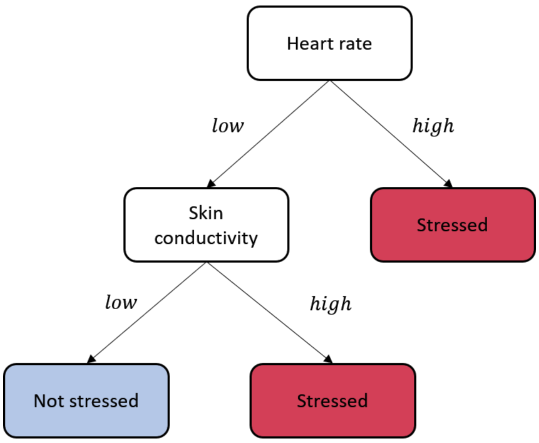

Figure 9. Toy Decision Tree to classify whether a person is stressed or not based on heart rate and skin conductivity.

Growing a Decision Tree works by recursively partitioning the feature space so that instances from the same class are grouped together [54]. Again let $X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n\}$

with $\boldsymbol{x}_i \in \mathbb{R}^d$ be a set of training samples and $Y = \{y_1, y_2, \dots, y_n\}$ the corresponding labels with $y_i \in \{0, 1\}$. The first step is to create a tree with one node (root) and quantify for each feature f_i with $i \in \{1, \dots, d\}$ the quality of a split $\theta(\mathrm{f_i}, \mathrm{t_m})$ based on this feature with a threshold t of split m. For this thesis this is done using the *entropy* (information gain) of a split[4]. For a binary classification problem, let

$$p_k = \frac{|y = k|}{|Y|} \tag{18}$$

be the proportion of samples belonging to class k regarding all samples of the training set, with $k \in \{0, 1\}$. The entropy H of a split Q_m is then defined as

$$H(Q_m) = -\sum_{i=1}^{k} p_k log(p_k) \tag{19}$$

so that the expected information gain (change in entropy) can be expressed as

$$IG(Q_m, \theta) = H\big(Q_{parent}\big) - \sum_{i=1}^{k} p_k\, H\big(Q_{child,k}\big) \tag{20}$$

with $H\big(Q_{parent}\big)$ and $H(Q_{child})$ being the information gain of the parent and the child node, respectively [54], [55]. The split is then made for the best candidate split

$$\theta^* = \mathrm{argmax}_\theta\, IG(Q_m, \theta)\ . \tag{21}$$

This process is repeated until the maximum depth is reached. The maximum depth for this thesis has been determined by cross-validation. Again the implementation of Python's sci-kit learn has been used [53]. Decision Trees come with the advantages of including feature selection and being easy to interpret. However, they are known to produce completely different trees with small variations in data [54]. Hence, interpretation should be made with caution as this thesis dataset only contains few data samples.

[4] Documentation sci-kit learn: https://scikit-learn.org/stable/modules/tree.html

3.5 Feature Selection

Feature selection aims at deciding which of the initial features to keep and which to discard. A brute-force approach of n features would need an evaluation of 2^n feature subsets. Even for a small feature set of size 10, this would lead to training and evaluating $2^{10} = 1024$ different models. As this is computationally costly, a brute force approach is not applicable. Therefore, various algorithms were developed to tackle the problem of finding the best features for prediction. In this thesis work, four algorithms have been applied to select the most predictive features.

As a pre-processing step all features were standardized to have a mean of 0 and a standard deviation of 1. This was done by using the mean μ and the standard deviation σ of each feature, to transform each datapoint X as follows:

$$z = \frac{X - \mu}{\sigma} \tag{22}$$

This was done for two reasons. Firstly, it significantly speeds up processing time, and secondly, was found to increase prediction accuracy [4], [28]. As Decision Trees are robust against scaling, data for this classifier has not been preprocessed.

Prior applying the feature selection algorithms to the thesis data, each algorithm has been tested on a benchmark dataset, to demonstrate that the algorithms can reduce the number of features without significantly reducing the predictive power. As the thesis problem is a binary classification problem, the results are verified using the artificial *madelon* dataset[5], which was part of the NIPS 2003 Feature Selection Challenge [35]. This balanced dataset consists of 2600 samples, 500 continuous features, and two discrete target classes. It should be noted that the general goal of feature selection algorithms is to reduce the number of features without losing predictive power. Increasing the accuracy is desirable and often the case, but not self-evident. All feature selection algorithms have been implemented in Python, and partially, to increase performance, in Cython. In the following the four feature selection algorithms are further described.

[5] https://www.openml.org/d/1485

3.5.1 Correlation-based

The correlation-based feature selection algorithm (CFS) according to Hall (2000) is a filter method, and therefore independent of the final prediction model. It evaluates the merit of a feature subset by considering the average predictive power of the features regarding the classification label, as well as the inter-correlation among the features in the subset. The goal is to find a feature subset with low feature-feature correlation, to avoid redundancy, and high feature-class correlation to keep only predictive ones [36]. The heuristic merit of a feature subset s with k features is defined as:

$$Merit_s = \frac{k \cdot \overline{r_{cf}}}{\sqrt{k + k \cdot (k-1) \cdot \overline{r_{ff}}}} \tag{23}$$

where $\overline{r_{ff}}$ is the average feature-feature inter-correlation and $\overline{r_{cf}}$ is the average feature-class correlation. Due to categorical labels of the discussed classification problem, the feature-class correlation is calculated using the point-biserial correlation coefficient. The features are all continuous and hence, the standard Pearson's correlation coefficient was used to calculate the feature-feature correlation. The average was computed using the absolute values of all means.

Given this heuristic, the correlation-based feature selection algorithm uses a best first search algorithm to search the feature subspace. According to Hall (2000) this yielded slightly better results than a hill climbing search. For this thesis the search algorithm uses a priority queue as data structure, where each inserted item has a priority associated with it, and the item with the highest priority is returned at request. If an item is already in the queue with a smaller priority, its priority gets updated. Each item of the queue is a feature subset, and the associated priority is the merit of this subset.

The search algorithm first starts with an empty feature subset s_0 and evaluates for each feature the merit of being added to the empty set. For this step the feature-feature correlation can be neglected, as the denominator of equation (23) can be simplified to one, due to $k = 1$. Therefore, only the feature-class correlation is considered, and the feature with the highest merit is added to the subset, so that s_1 contains this feature. In the next step again all features, except for the one already added, are evaluated, and the feature that forms the best subset combined with the previously added one, is kept. This process is iterative, and whenever an expansion of features yields no improvement, the algorithm drops back to the next best unexpanded subset. Without a limitation this algorithm searches the whole feature subset space, hence, the number of backtracks is limited to n. If the algorithm checks the expanded features n times without improving the merit, it terminates and returns the feature subset that achieved the highest merit up to this point.

3.5.2 Relief-based

The Relief-based feature selection algorithm is a filter method and therefore independent of the final classification model [33]. In contrast to the correlation-based feature selection algorithm, the Relief-based algorithm weights and ranks individual features instead of subsets. Hence, it is possible to either define the subset size manually or include all features that exceed a certain relevance threshold. One key advantage is that it is a statistical method, avoiding any heuristic search, which is why it only requires linear time [56]. The algorithm estimates the quality or relevance of a feature regarding a target by calculating a proxy statistic. These are called *feature weights* with $W[f_i]$ denoting the weight or relevance of feature f_i. Let X be a dataset with n samples, m numerical features, and corresponding binary target values $y \in \{0,1\}$. At start the Relief-based algorithm initializes all feature weights with $W[f_i] \coloneqq 0$ for all $i \in \{1, \dots, m\}$. Then it iteratively updates the weights for each feature based on differences to close instances. In each cycle the algorithm randomly draws, without replacement, an instance $\boldsymbol{x} = \{x_1, \dots, x_m\} \in X$ with the corresponding target value $y \in \{0,1\}$ and determines the nearest neighbor with the same target, called *nearest hit* (H), and the nearest neighbor with the opposite target, called *nearest miss* (M). The term *nearest* refers to the distance of two instances in feature space and is determined by

$$\mathrm{d}(\mathbf{x}, \mathbf{x}') = \sum_{\mathrm{i}=1}^{\mathrm{m}} |x_i - x_i'| \tag{24}$$

which is called Manhattan distance [33]. In the original Relief algorithm the Euclidean distance was used [56], but as no significant differences were found, the Manhattan distance has become standard [33]. Based on the three instances x, M, and H, the weights of each feature are updated. When comparing the feature values of the nearest miss M and the randomly drawn instance $\boldsymbol{x}$, features whose values differ can be considered as informative of outcome. Hence, the weights of these features are increased. On the contrary, feature value differences between $\boldsymbol{x}$ and H, imply that these features might not be discriminative, and the weights are decreased. The weight update magnitude for continuous features is calculated using the function

$$diff(f, \boldsymbol{x}, \boldsymbol{x}') = \frac{|value(f, \boldsymbol{x}) - value(f, \boldsymbol{x}')|}{max(f) - min(f)} \tag{25}$$

with f being the feature under evaluation, $\boldsymbol{x}, \boldsymbol{x}' \in X$ instances, $max(f)$ and $min(f)$ the maximum and minimum of feature f over all instances, respectively, and $value(f, \boldsymbol{x})$ the value of feature f of the instance $\boldsymbol{x}$. The normalization with the difference of the feature maximum and minimum ensures that the update falls between 0 and 1. The weights are then updated using

$$W[f_i] \coloneqq W[f_i] - \frac{diff(f_i, \boldsymbol{x}, H)}{k} + \frac{diff(f_i, \boldsymbol{x}, M)}{k} \tag{26}$$

with k being the number of iterations, or number of drawn training instances out of n, used to update the weights. This normalization ensures that the final weights are in the interval $[-1, +1]$ [33]. An example of one cycle with continuous feature values can be seen in Figure 10. After k cycles the algorithm returns a weight vector W which assigns a relevance or quality to each feature. Negative values or values close to zero indicate irrelevance of a feature, whereas positive feature suggest that a feature is relevant [33], [56]. The pseudo-code in algorithm 1 further illustrates the procedure.

The only parameter for this algorithm is the number of training iterations, or number of draws of instances. Finally, the algorithm allows to select the number of features manually. Several extensions of the Relief algorithm have been proposed. An example is *ReliefF* [57], which considers multiple nearest neighbors rather than just one, and can handle incomplete data, as well as multi-class problems. However, as this thesis data set is neither incomplete, nor multi-class, the original algorithm has been applied.

Algorithm 1 Pseudo-code for the Relief algorithm

Initialize all feature weights $W[f] \coloneqq 0$
repeat $\boldsymbol{k}$ **times**
  randomly draw an instance $\boldsymbol{x}$
  find nearest hit H
  find nearest miss M
  for f **in** features **do**
    $\boldsymbol{W[f] \coloneqq W[f] - diff(f, x, H)/k + diff(f, x, M)/k}$
return vector W of feature relevance scores

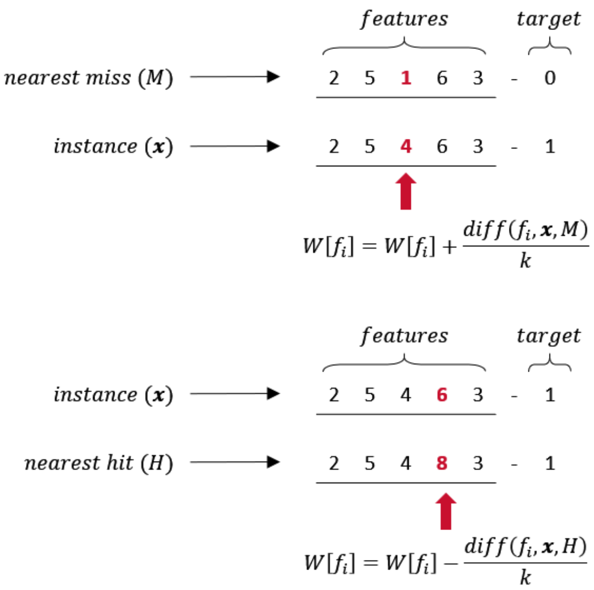

Figure 10. The feature weight updates for a randomly drawn instance $\boldsymbol{x}$ based on the nearest hit (H) and the nearest miss (M).

3.5.3 Genetic Algorithm

The genetic algorithm is a wrapper feature selection method and has its roots in evolution, where in each generation only the fittest survive [34], [37]. In terms of a feature selection method, each generation consists of a pool of hypotheses, with each representing a feature subset. These hypotheses are evaluated based on a fitness function, which for this thesis is a binary classification model (SVM). A bitstring is used to represent each feature subset, consisting of concatenated zeros and ones, with length equal to the feature set length. Hence, a one in the bitstring at position i shows that the feature f_i at position i in the feature set S is included, and in case of a zero, it is not. To further clarify this approach, assume a feature set with four features $S := f_1, f_2, f_3, f_4$ and a bitstring $B := 0110$ representing a feature subset s. The bitstring acts like a mask for the feature set S, and thus, the feature subset s includes only feature f_2 and f_3, as only B_2 and B_3 are equal to one. This example is illustrated in Figure 11.

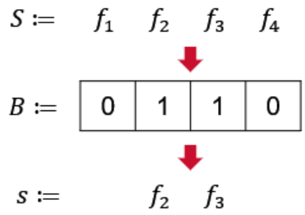

Figure 11. Feature subset representation in the genetic algorithm with a bitmask B,feature set S and subset s.

Three main operators are used for the evolution of populations: *Crossover*, *Mutation*, and *Selection*. Crossover mixes the bitstrings from two parents to assemble them to two offsprings. This is done by splitting the bitstrings of both parents at position i and combining the parts to create new compositions. Let parent 1 have the bitstring 011011 and parent 2 the bitstring 101010. Splitting the parents at position 3 yields two new offsprings having the bitstrings 011010 and 101011, respectively. This mechanism follows the idea of inheriting possibly successful parts of solutions from the parents, so that their combination might enhance the subset fitness [37]. For the implementation of this algorithm mates were selected randomly and crossover of two parents was performed with probability p_c.

The second main operator is mutation, which represents random changes [34], [37]. With probability p_m an individual of the population mutates and each bit of the n bits has probability $1/n$ to be flipped, either from 1 to 0, or vice versa. For example, mutating the bitstring 0101 could yield 1101 where only the bit at position one has been changed. This operator avoids getting stuck in a population and ensures that the whole search space is investigated by introducing a chance to reach all combinations of the solution space [37], [58].

The last operator is selection, which can be separated in two parts. Firstly, it evaluates the fitness of each individual of a population, and secondly, selects k individuals as new parents. The evaluation is based on the final model used for classification; hence, this feature selection method is model-dependent. This work used a Support Vector Machine for binary classification as model which is further described in section 3.4.1.

As the fitness of each offspring is solely based on its prediction score, there is no restriction for the algorithm to take as much features as possible. To overcome the problem of having large feature subsets a regularization has been introduced, which regularizes the score of an individual as follows. Let S be the feature set and $s_i \subset S$ a feature subset, with $|S|$ or $|s_i|$ the length of the sets, respectively. The function $F(s_i)$ associates each subset with a fitness value, using the accuracy it achieved in the classification model and the regularization term which can be written as:

$$F(s_i) = Accuracy(s_i) - \lambda \frac{|s_i|}{|S|} \tag{27}$$

In equation (27), the scalar λ acts as regularization rate, with higher values penalizing larger feature subsets more. Assuming a high regularization rate, short feature subsets that achieve a high accuracy in the classification model, get the highest fitness score, whereas large feature subsets or feature subsets that perform poorly in the classification, get a low fitness score. To avoid negative values the fitness was computed as $max\big(0, F(s_i)\big)$.

In the beginning of the algorithm the population is initialized randomly. Entering a loop with i iterations, first, the fitness for each individual is computed and the k best individuals are selected to be the new parents of that generation. Thereafter the population undergoes crossover, for randomly chosen pairs, and mutation. After i iterations, the algorithm selects j individuals from the final population as the resulting feature subsets. The whole procedure can be seen in algorithm (2).

Algorithm 2 Basic Genetic Algorithm

```
initialize population
repeat i times
    for each individual do
        fitness computation
    selection
    for each pair do
        crossover
    for each individual do
        mutation
select final individuals
```

3.5.4 Embedded Method

The last method that has been applied is the embedded feature selection method, where feature selection is already integrated in the classification algorithm itself. The algorithm of choice is a Decision Tree classifier, which has already been described in section 3.4.3. Note, that for this algorithm features were not standardized, as Decision Trees are robust against scaling. Furthermore, as for this algorithm the sci-kit learn implementation has been used, the algorithm has not been applied to the benchmark dataset.

4. Results

4.1 Benchmark

For all benchmark models a SVM with radial-basis kernel was used and the parameters C and γ set to 100 and 0.01, respectively. Every reported accuracy is the mean of a 10-fold cross-validation SVM. However, for randomly generated subsets only a train-test split of 80/20 has been used to evaluate the model. Running a SVM with all 500 features as input yielded a mean accuracy of 59.85% within 43.8 seconds.

Correlation-based Feature Selection. The correlation-based feature selection algorithm applied to the benchmark dataset with a maximum number of backtracks equal to ten, reduced the number of features from 500 to 48, which is a reduction of approximately 90%. Running a SVM with this feature subset resulted in a mean accuracy of 61.58% in 10.07 seconds. Hence, the above-described algorithm successfully reduced the number of features, which decreased model runtime, and was even able to increase accuracy. To prove that this increase does not come by chance, the SVM model got trained with 100 different, randomly generated feature subsets with same size as the determined feature subset. For this, the dataset was split into 80% and 20%, for training and testing, respectively. The resulting accuracies can be seen in Figure 12 and indicate that the selection of features does not come by chance. Furthermore, it significantly improved accuracy.

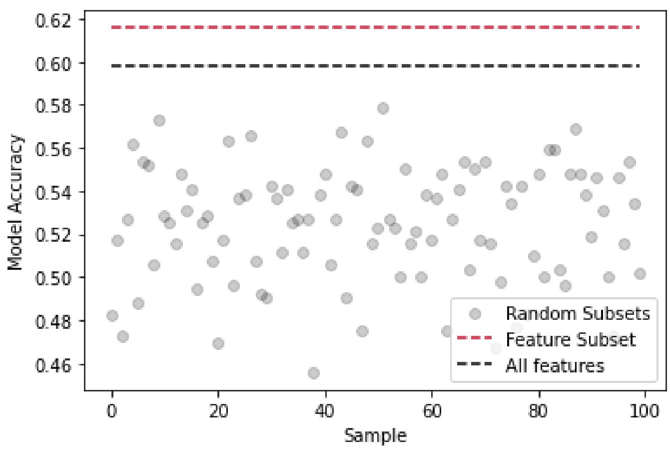

Figure 12. The accuracies of 100 randomly chosen feature subsets with the same size as the subset determined by the correlation-based algorithm and their corresponding accuracies, compared to the accuracies of the model trained with all features (black dashed line) and the model trained with the selected feature subset (red dashed line), respectively.

Relief-based Feature Selection. Running the Relief-algorithm on the benchmark dataset returned an array of weights for each feature. Hence, the subset size needed to be determined manually. For this, the features were ordered based on their weight and 100 subsets created by gradually adding the features, so that the first subset contained just the feature with largest weight, and the last subset contained the 100 features with largest weights. Results are displayed in Figure 13.

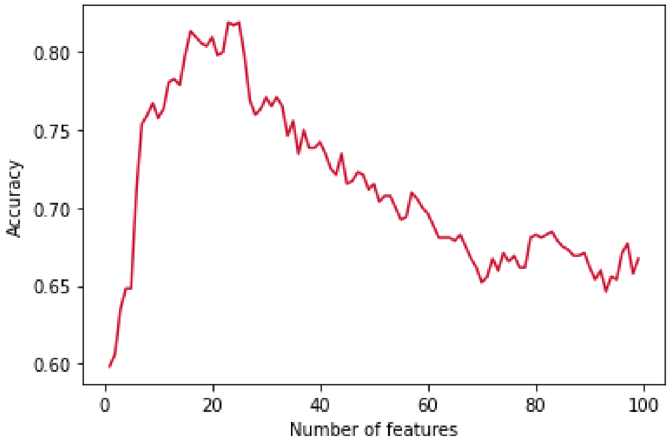

Figure 13. SVM accuracies for 100 subsets with incrementally increasing subset size based on the features according to their Relief-based weights.

Including 24 features with largest weight yielded the best result and hence, the final benchmark model was trained with that subset. The Relief-based feature selection thus significantly reduced the number of features from 500 to 24 (~95% reduction) and a 10-fold cross-validation SVM achieved a mean accuracy of 81.46%. Model run time was 12.7 seconds. Hence, it can be concluded that the Relief-based feature selection successfully reduced the number of features and increased model accuracy of approximately 21%. To exclude that any randomly selected feature set of same size reaches this result, 100 subsets with 24 features were randomly drawn and evaluated. The accuracies can be seen in Figure 14 and indicate that the achieved accuracy of the Relief-based feature selection subset did not come by chance.

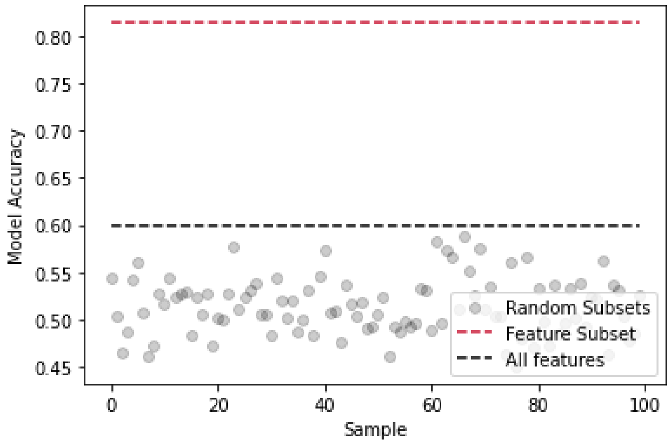

Figure 14. The accuracies of 100 feature subsets, each consisting of 24 randomly chosen features, compared to the accuracies of a model considering all features (black dashed line) and the subset selected by the Relief-based algorithm (red dashed line), respectively.

Genetic Algorithm. The settings for the Genetic Algorithm applied on the benchmark dataset were as follows: initial population 16, number of individual selections per epoch 8, crossover rate $p_c = 0.9$, mutation rate $p_m = 0.5$, and 300 epochs. Without applying regularization (setting $\lambda = 0$ in equation (27)), the Genetic Algorithm reduced the number of features from 500 to 235, which is a reduction of about 53%. The mean 10-fold cross-validation accuracy for the subset selected by the Genetic Algorithm without regularization was 61.81% and training the model took 16.1 seconds. Thus, the algorithm reduced computational complexity and slightly increased accuracy, however, not significantly. Nonetheless, 100 randomly selected feature subsets with same size could not approach this accuracy, as illustrated in Figure 15. This indicates that the increase in accuracy could not be achieved with random feature subset constellations of size 235.

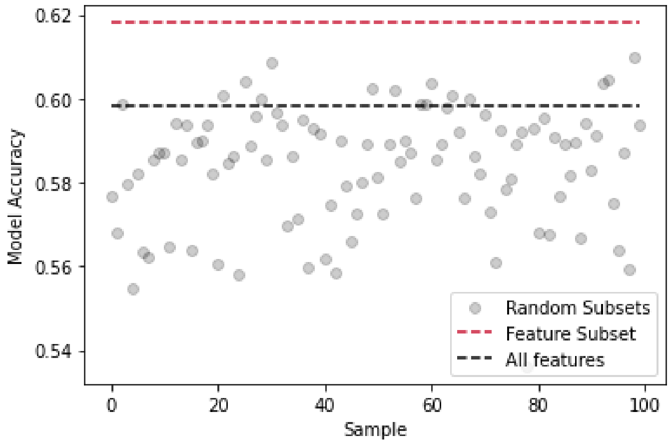

Figure 15. The accuracies of 100 feature subsets each having different, randomly chosen feature constellations of 235 features. These are compared to the accuracies of a model considering all features (black dashed line) and the subset selected by the Genetic Algorithm without regularization (red dashed line), respectively.

However, the results of the Genetic Algorithm without regularization emphasize the need for a subset size penalty as described in equation (27). Therefore, the same analysis was performed with the regularization parameter λ set to one. This penalizes the fitness of a subset with a magnitude of subset length divided by the total number of available features. Applying the Genetic Algorithm with this regularization reduced the number of features from 500 to 171, or about 66%. Compared to the results without regularization, this is a further decrease of approximately 13%. The penalizing model reached a mean 10-fold cross-validation accuracy of 61.8%, with a training time of 12.2 seconds. Figure 16. shows that 100 randomly generated subsets, each consisting of 171 randomly chosen features, could not achieve an equivalent or greater accuracy than the subset selected by the Genetic Algorithm with regularization. Hence, it can be concluded that the Genetic Algorithm, with and without regularization, is able to significantly reduce the input features without losing predictive power. Furthermore, the added feature subset size regularization helps to avoid large feature subsets.

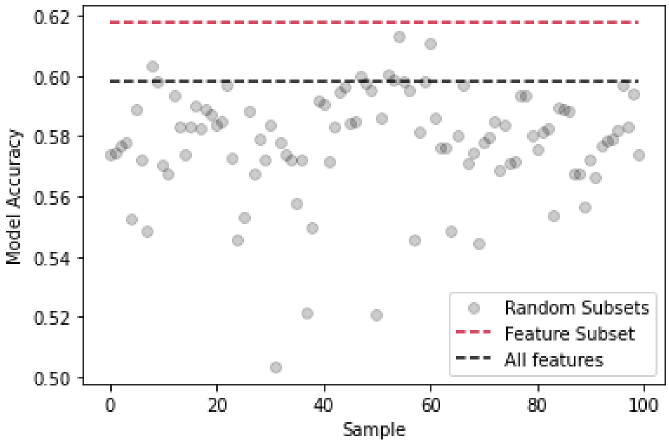

Figure 16. The accuracies of 100 feature subsets each having different, randomly chosen feature constellations of 171 features. These are compared to the accuracies of a model considering all features (black dashed line) and the subset selected by the Genetic Algorithm with regularization $\lambda = 1$ (red dashed line), respectively.

Summary. The previous results indicate that all three feature selection methods were able to reduce the number of features and thus, computational complexity. Furthermore, they maintained or even increased predictive power. While the Genetic Algorithm (with and without regularization) and the Correlation-based feature selection did not increase accuracy significantly, the Relief-based algorithm was able to increase the mean accuracy from 59.85% (all features) up to 81.46%. For this it only used 4.8% of all features. Hence, it can be concluded that the implemented feature selection methods are a proper tool to reduce model complexity and maintain or increase accuracy by discarding irrelevant, noisy features. In Table 2 all feature selection results are summarized.

Table 2. Comparison of the results of the feature selection algorithms applied on the benchmark dataset. GA = Genetic Algorithm, CFS = Correlation-based feature selection, Relief = Relief-based feature selection. The reported average accuracies are the result of a 10-fold cross-validation Support Vector Machine model. Best results are displayed in bold face.

Feature Set	Set size		Time	Accuracy
	Number	Percent	(seconds)	(mean)
All features	500	100.0 %	43.8	59.85 %
GA normal	235	47.0 %	16.1	61.81 %
GA regularized	171	34.2 %	12.2	61.80 %
CFS	48	9.6 %	10.0	61.58 %
Relief	24	4.8 %	12.7	**81.46 %**

4.2 Galvanic Skin Response

The galvanic skin response (GSR) was recorded from 17 participants, however, due to missing data two subjects had to be excluded. From the remaining 15 subjects 8 were male, mean age was 19.87 years (standard deviation: 1.78). The descriptive statistics per factor *state* (normal vs. relaxed) and *eyes* (open vs. closed) are presented in Table 3.

Table 3. For each factor (State and Eyes), the mean skin conductivity in micro siemens.

	State	
	Normal	Relaxed
Eyes		
Open	4.084	3.142
Closed	2.822	2.776

A *Shapiro-Wilk* test revealed that the normality assumption was violated for both factors (*State* and *Eyes*). The distribution per combination of factors can be seen in Figure 17. Since the repeated-measures ANOVA is robust to violations of the normal distribution assumption, an analysis was still performed. There was no significant interaction effect between the two factors *State* and *Eyes* ($F(1,14) = 3.61, p = 0.08, \eta_p^2 = 0.21$), nor a significant main effect of *State* ($F(1,14) = 1.20, p = 0.29, \eta_p^2 = 0.08$). Only the null hypothesis of the factor *Eyes* is rejected based on the sample ($F(1,14) = 8.93, p < 0.01, \eta_p^2 = 0.39$).

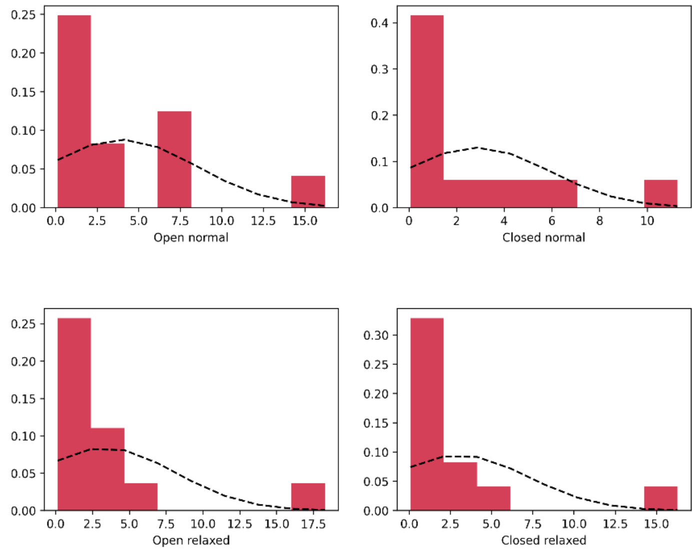

Figure 17. The distribution of the dependent variable *GSR* for each combination of the two factors *State* (normal vs. relaxed) and *Eyes* (open vs. closed).

4.3 Feature Selection

At the time when the feature selection analysis was performed, data from only 23 subjects were available. Due to missing data one further subject had to be excluded. Mean age of the remaining 22 subjects (6 female) was 22.64 years (standard deviation: 4.86). For the application of the feature selection methods no differences were made with regards to open or closed eyes. Therefore, for each participant each eye condition was treated as independent sample. Hence, the dataset contained 88 samples, 44 per class.

Correlation-based Feature Selection. For the correlation-based feature selection algorithm applied on the thesis data the maximum number of backtracks was set to 400 and reduced the number of features from 47 to 7. The selected features for the best subset were:

- Mean Tidal Volume
- Mean Pulmonal Ribcage Volume
- Mean Abdominal Volume
- Frequency of mean values of the Px spectrum
- Mean Abdominal Ribcage Volume
- Maximum peak of the Px spectrum
- Standard deviation of Abdominal Ribcage Volume

Five of the seven selected features were breathing variables, whereas only two stabilometry features were included, namely the maximum peak of the Px spectrum, and the frequency of mean values of the Px spectrum.

Relief-based Feature Selection. With 1000 iterations, the Relief-based Algorithm returned an array of feature relevance weights. To determine the best feature subset size multiple 10-fold cross-validation SVMs were run, and their accuracies used as heuristic. This revealed that the best subset only includes 5 of all 47 features. The features ordered according to their weights from top to bottom were as follows:

- Mean Pulmonal Ribcage Volume
- Mean Tidal Volume
- Mean Abdominal Volume
- Mean Abdominal Ribcage Volume
- Difference Mean Abdominal Volume and Pulmonal Ribcage Volume

The best subset of the Relief-based feature selection algorithm only found breathing variables to be predictive of the state. The first stabilometry feature appeared on the ranked feature list on place eight.

Genetic Algorithm. The Genetic Algorithm was run for 300 epochs with an initial population size of 32, a mutation rate of $p_m = 0.5$, a crossover rate of $p_c = 0.9$, a regularization of $\lambda = 0.6$, 10-fold cross-validation, and 16 selection per epoch. From 47 features the Genetic Algorithm has selected the following 5:

- Mean Tidal Volume
- Frequency of mean values of the Px spectrum
- Equivalent area
- Speed
- Geographic Area

In contrast to the Relief-based and the Correlation-based feature selection, this subset contains mainly stabilometry measures and only the mean tidal volume as breathing variable.

Decision Tree. The Decision Tree with a maximum depth of four identified the following five features:

- Mean Tidal Volume
- Frequency of mean values of the Px spectrum
- Frequency of mean values of the Py spectrum
- Transversal range
- Minimum radius

Four of the five selected features are stabilometry variables, only one is a breathing variable.

Summary. Taken together all four feature selection methods included the mean tidal volume as one of the predictors in their feature subsets. While the Correlation-based and Relief-based feature selection methods mostly identified breathing variables as predictive, it is the opposite for the Decision Tree and the Genetic Algorithm, which identified more stabilometry measures. For a further analysis the top five feature subsets of the Genetic Algorithm, the Relief-based, and the Correlation-based feature selection algorithm have been identified, and the number of appearances per feature counted. For the Genetic Algorithm a final selection of five was used. As the Correlation-based feature selection algorithm was implemented by using a Priority Queue, the top five subsets with highest merit have been selected, and for the Relief-based algorithm, feature subsets of size four to eight were created by incrementally adding the next best feature. In the resulting list of features over all subsets and methods the appearance of each feature was counted. This approach behaves like a voting mechanism, where each word appearance votes for a certain feature. The top ten features of this voting are displayed in Figure 18 indicate that breathing measures are the most appropriate predictors, as the volume of each compartment (Pulmonal Ribcage, Abdominal Ribcage, and Abdominal) and the tidal volume are among the top four features. However, on the fifth place the first stabilometry measure with the transversal range is present. It should be mentioned that for this voting the Decision Tree Algorithm has been neglected, as it is not able to produce several subsets. Due to that the voting may be in favor of the breathing variables, as the Relief-based and Correlation-based method mostly considered breathing measurements as predictive.

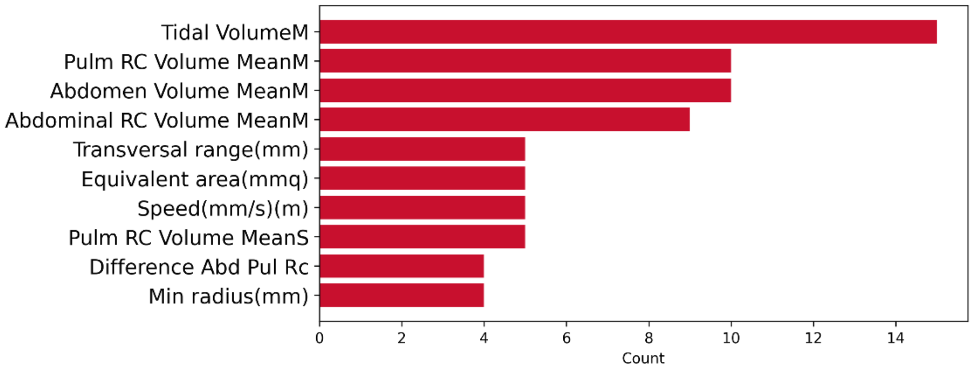

Figure 18. The top ten features that got the most votes of the best five feature subset selected by the Genetic Algorithm, the Correlation-based method, and the Relief-based method. Abbreviations: *MeanM* = mean, *MeanS* = standard deviation, *Difference Abd Pul Rc* = Abdominal Volume minus Pulmonal Ribcage Volume.

4.4 Classification

For the predictive modelling step three different data constellations have been used. Firstly, the condition eyes open versus eyes closed has been treated as independent samples, so that each participant appeared twice in each state (normal and relaxed). Secondly, feature selection and classification were performed for two data subsets containing only open and closed eyes, respectively. Lastly, only stabilometry measures have been used for feature selection and classification to see if these are sufficient as input for the state prediction. For each of the three data constellations a *leave-one-subject-out* (LOSO) k-nearest neighbor classification model, as well as a LOSO Support Vector Machine have been trained. The parameters for the latter have been determined using a grid-search with 10-fold cross-validation for $C \in \{10^x \mid x \in [-1, 6]\}$ and $\gamma \in \{10^x \mid x \in [-5, 2]\}$.

Eyes mixed. Treating the eyes open and eyes closed condition as different samples resulted in a dataset of 96 samples, with 48 of each class. One subject had to be excluded due to missing data. The accuracies per feature selection method and classification model can be found in Table 4. The highest accuracy with 85.42% has been achieved with the SVM model and the feature subset selected by the Genetic Algorithm. All selected feature subsets performed better than a model using all 47 features. Hence, the feature selection methods were able to significantly reduce the number of features and increase accuracy. For the KNN model the feature subsets increased accuracy with at least 7%. The subsets fed to the SVM also increased accuracy, however, not always significantly. Results are further illustrated in Figure 19.

Table 4. Feature subset size and *leave-one-subject-out* mean accuracy per feature selection algorithm and classification model compared to a model with all features as input variables. Feature Selection methods: CFS = Correlation-based feature selection, GA = Genetic Algorithm, Relief = Relief-based Algorithm. Models: KNN = k-nearest neighbor, SVM = Support Vector Machine. The best accuracies per model are displayed in bold. Note that for this feature selection and classification the state of the eyes has been neglected and treated as different subjects. Hence, 88 samples were used.

Feature Set	Set size		Mean Accuracy (%)	
	Number	Percent	KNN	SVM
All features	47	100.0 %	72.92	80.21
CFS	7	14.9 %	**84.38**	82.29
GA	5	10.6 %	80.21	**85.42**
Relief	5	10.6 %	83.33	83.33

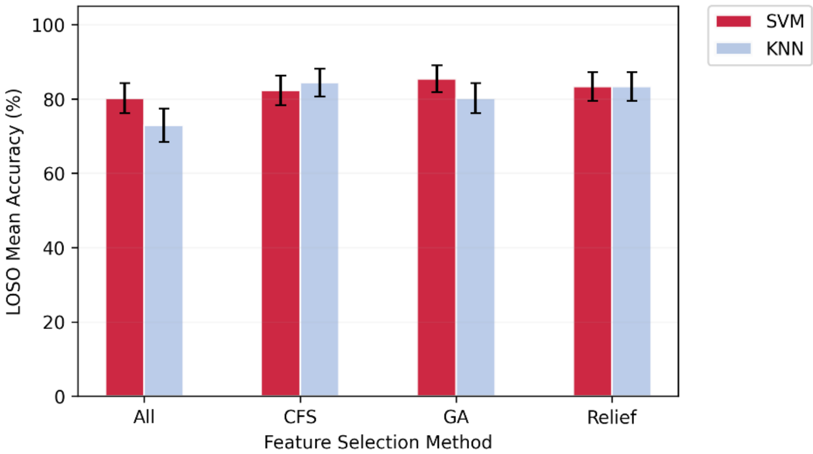

Figure 19. Mean *leave-one-subject-out* accuracies per feature selection method and classification model. Error bars represent the standard error of the mean. Feature Selection methods: **CFS** = Correlation-based feature selection, **GA** = Genetic Algorithm, **Relief** = Relief-based Algorithm. Models: **KNN** = k-nearest neighbor, **SVM** = Support Vector Machine. Note that for this feature selection and classification the state of the eyes has been neglected and treated as different subjects. Hence, the dataset consisted of 96 samples.

For the Decision Tree the data has not been standardized as Decision Trees are robust against scaling. Running the algorithm on the dataset with a maximum depth of 5 yielded the tree displayed in Figure 20. The entropy indicates the information gain of splitting a node. The tree only selected one breathing variable and four stability measures. The first split of the tree is based on the mean tidal volume and samples that have a mean tidal volume smaller than 0.725, are correctly assigned to the normal class (Stress). Thereafter, samples are split according to the mean frequency of the spectrum in the x direction, with low values indicating the relaxed condition. Then again, the mean tidal volume is used to split. Samples with higher values are considered to be in the relaxed condition and smaller values are further split on the transversal range. A large transversal range suggests the normal state, whereas smaller values are split again. The LOSO mean accuracy was 78.12%. However, as Decision Trees are highly dependent on the samples and the number of the samples, interpretations should be made with caution.

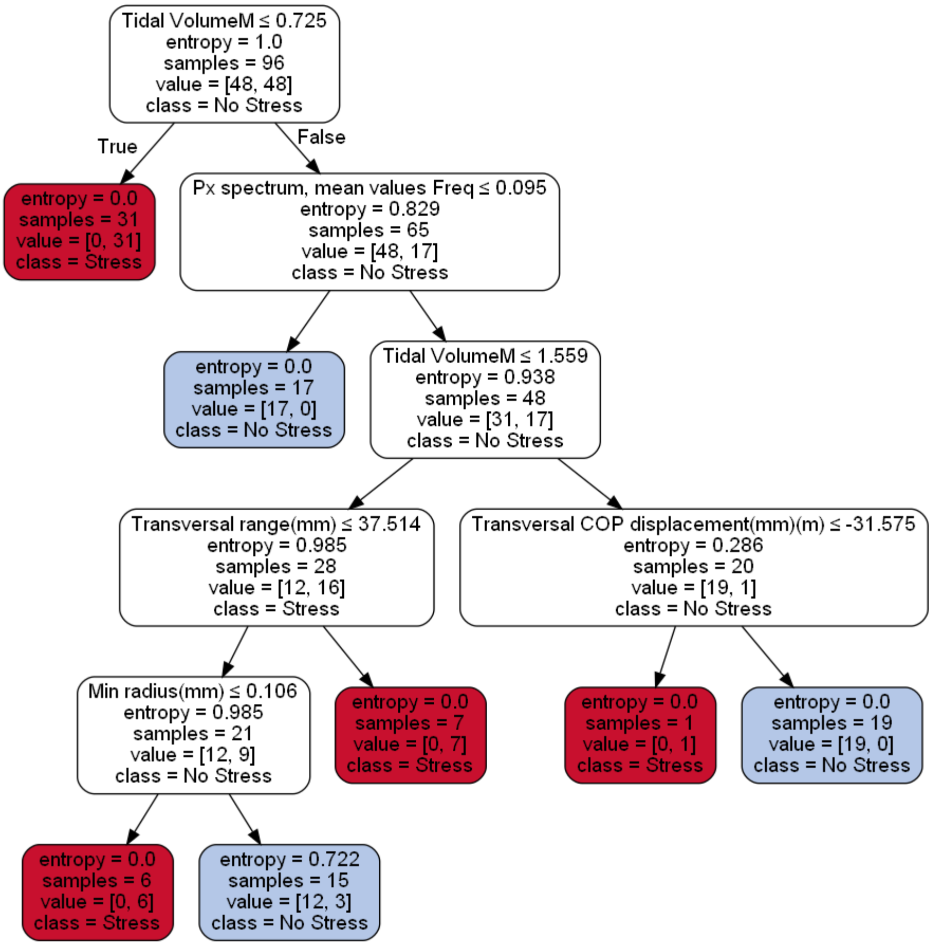

Figure 20. Decision Tree for the whole dataset treating open and closed eyes condition as individual samples. *Stress* refers to the normal quiet standing state and *No stress* to the relaxed quiet standing after participants performed muscle relaxation and deep breathing exercises.

The top three features according to both, the feature voting of section 4.3, as well as the Relief-based feature selection weights, are the mean tidal volume, the mean pulmonal ribcage volume, and the mean abdominal volume. In Figure 21 these three variables are plotted in a three-dimensional space for each state of quite standing (normal vs. relaxed) and each state of the eyes (open vs. closed). It seems that if all three variables are close to the origin (0,0,0) the state is very likely the normal quite standing, regardless of the state of the eyes. Most of the datapoints with very high values on each of the three dimensions can very likely be assigned to the relaxed state. It even seems like data in this three-dimensional space may be linearly separable by a hyperplane, of course, at cost of misclassifications.

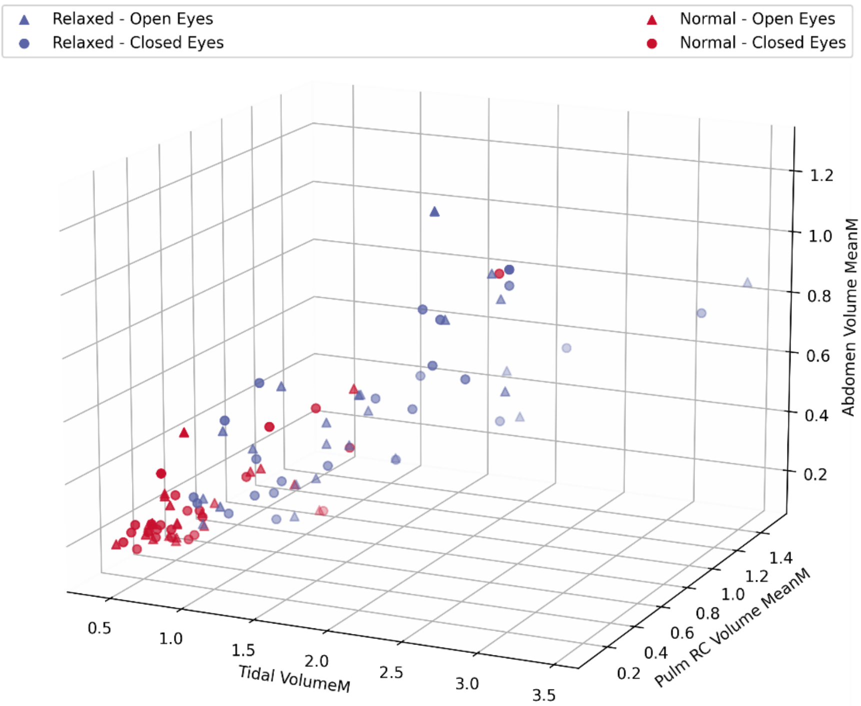

Figure 21. 3-dimensional representation of the three best features according to the weights of the Relief-based algorithm, as well as the voting results of section 4.3. Blue represents the relaxed quiet standing, whereas red represents the normal quiet standing. Triangles are standing for open eyes, circles for closed eyes.

Eyes separated. For this feature selection and classification two data-subsets have been created, one only including open eyes trials, and one only including closed eyes trials. Again, due to incomplete data on subject had to be excluded. Hence, each of the subsets contained 48 samples, 24 for each class. Results are displayed in Table 5. The highest accuracy for the open eyes condition has been achieved with the subset selected by the

Genetic Algorithm and a SVM with a mean LOSO accuracy of 95.83%. For the closed eyes data-subset the highest mean LOSO accuracy was 85.42% with the Correlation-based and Genetic Algorithm, both used with a SVM. Each of the feature selection methods reduced the number of input variables significantly. In the open eyes condition every model trained with a feature subset performed better than a model including all features, with at least 6% increase for the KNN and at least 2% increase for the SVM model. In the closed eyes condition all feature subsets increased accuracy of at least 2% for the SVM and at least 16% for the KNN model. It can therefore be concluded that the feature selection method applied to both, the open eyes and the closed eyes data-subset reduced model complexity significantly, and increased model accuracy for all subsets. The mean LOSO accuracies for the open (top) and closed (bottom) eyes condition per feature selection algorithm and model are further illustrated in Figure 22.

Table 5. For two dataset-subsets, containing only open eyes or closed eyes samples, respectively, the feature subset size and *leave-one-subject-out* mean accuracy per feature selection algorithm and classification model compared to a model with all features as input variables. Feature Selection methods: CFS = Correlation-based feature selection, GA = Genetic Algorithm, Relief = Relief-based Algorithm. Models: KNN = k-nearest neighbor, SVM = Support Vector Machine. The best accuracies per model and condition are displayed in bold.

Feature Set	Set size		Mean Accuracy (%)	
	Number	Percent	KNN	SVM
Open Eyes				
All features	47	100.0 %	72.92	81.25
CFS	9	19.1 %	79.17	91.67
GA	7	14.9 %	**81.25**	**95.83**
Relief	6	12.7 %	**81.25**	83.33
Closed Eyes				
All features	47	100.0 %	62.50	77.08
CFS	8	17.0 %	79.17	**85.42**
GA	5	10.6 %	81.25	**85.42**
Relief	6	12.7 %	**83.33**	79.17

The feature subsets selected by the Genetic Algorithm that achieved an accuracy of 95.83% in the open eyes condition consisted of the following 7 features:

- Mean Tidal Volume
- Standard Deviation of the Longitudinal COP displacement

- Transversal Range
- Regression Angle
- Peak number
- Standard deviation of the Peak Amplitude
- Maximum peak of the Px spectrum

Hence, this high accuracy has been achieved with only one breathing variable and six stabilometry variables. However, further analysis without the breathing variable and just the remaining stabilometry measures only achieved a mean LOSO accuracy of 64.58% in the open eyes condition. The same applies for the subset selected by the Genetic Algorithm in the closed eyes condition where the subset contains the following features:

- Mean Pulmonal Ribcage Volume
- Minimum radius
- Mean Peak distance
- Frequency of mean values of the Px spectrum
- Frequency of maximum values of the Py spectrum

Creating a subset without the mean Pulmonal Ribcage Volume, which is the only breathing variable in this subset, results in a mean LOSO accuracy of 68.75%. This suggests that stabilometry measures alone may not be as predictive to detect the normal or relaxed stress, respectively, as if breathing variables were included. This is further analyzed in the following.

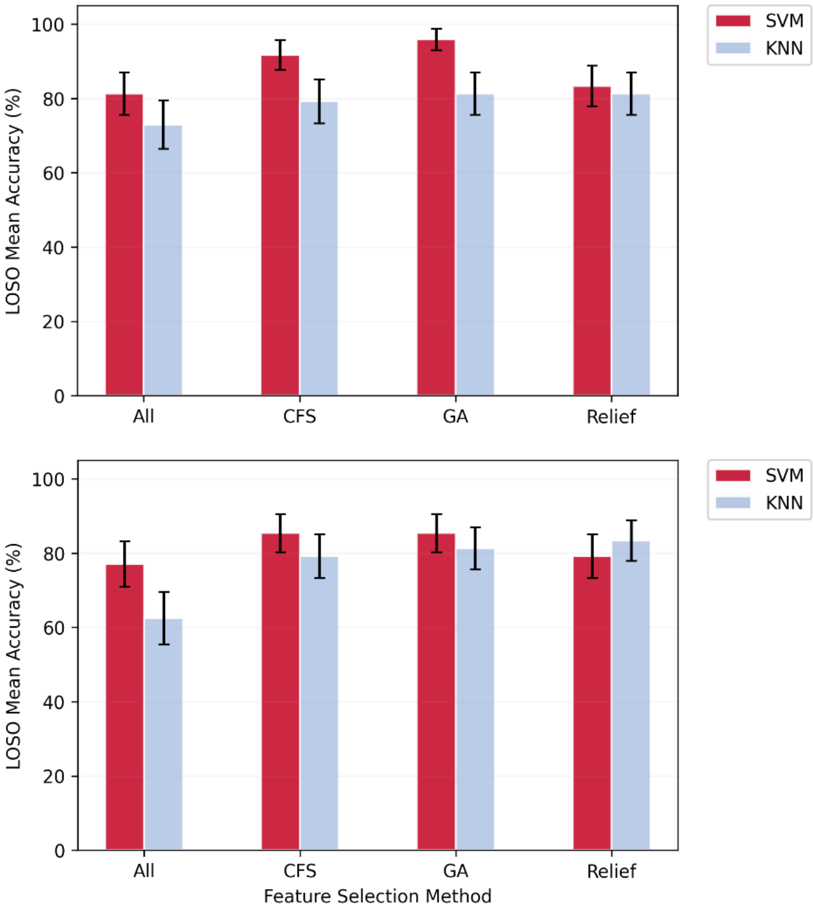

Figure 22. Mean *leave-one-subject-out* accuracies per feature selection method and classification model for an open eye (top) and closed eye (bottom) data-subset. Error bars represent the standard error of the mean. Feature Selection methods: **CFS** = Correlation-based feature selection, **GA** = Genetic Algorithm, **Relief** = Relief-based Algorithm. Models: **KNN** = k-nearest neighbor, **SVM** = Support Vector Machine. Each data-subset consists of 48 samples.

Stabilometry. Running the feature selection and classification with only stabilometry variables regardless of the state of the eyes yielded the results displayed in Figure 23. These indicate that stabilometry measures alone may not be sufficient to detect the normal or relaxed state as all LOSO mean accuracies are in the range of 40-60%. Due to the balanced dataset this is not better than guessing. Considering only a data-subset with open or closed eyes, yielded accuracies of 66.67% and 58.33%, respectively. These findings indicate that stabilometry variables alone may not be sufficient to predict the relaxed or normal state.

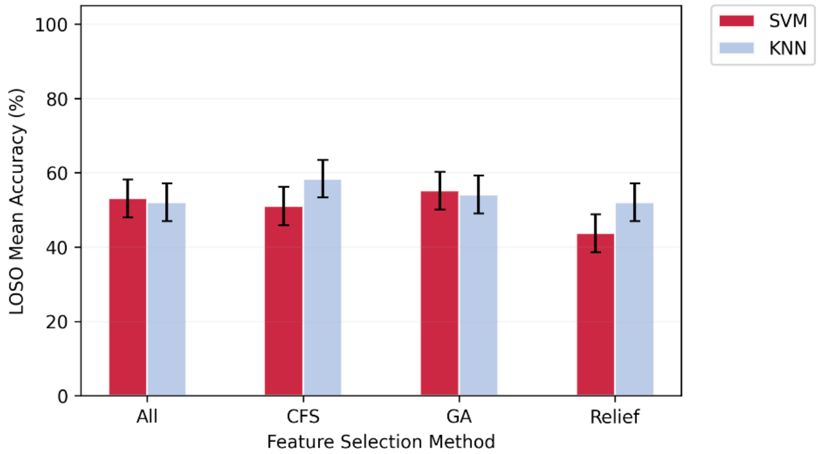

Figure 23. Mean *leave-one-subject-out* accuracies per feature selection method and classification model. Error bars represent the standard error of the mean. Feature Selection methods: **CFS** = Correlation-based feature selection, **GA** = Genetic Algorithm, **Relief** = Relief-based Algorithm. Models: **KNN** = k-nearest neighbor, **SVM** = Support Vector Machine. Note that for this feature selection and classification the state of the eyes has been neglected and treated as different subjects. Hence, the dataset consisted of 96 samples.

Exploratory Labelling. For the exploratory analysis the labels stress versus no stress were chosen based on the differences between mean abdominal and mean pulmonal ribcage volume in the relaxed state regardless of the state of the eyes. Since this dataset is unbalanced through this labelling anyway, all participants were included, even the one with missing data. This led to a dataset with 33 stressed and 16 not-stressed samples, respectively. Hence, an algorithm that always classifies a sample as stressed would achieve an accuracy of 67.35%. Input features were only stabilometry measures. Using all features resulted in a mean LOSO accuracy of 79.59% for the KNN and 73.47% for the SVM. A SVM trained with the subset selected by the Genetic Algorithm achieved a mean LOSO accuracy of 87.76% with the following 10 features:

- Mean of the Longitudinal COP displacement
- Standard deviation of the Radius
- Maximum Radius
- Longitudinal Range
- Equivalent Area
- Inertial Axis Y
- Mean Peak Amplitude
- Mean Peak Time

- Frequency of the Maximum Peak of the D Spectrum
- Frequency of mean values of the Px spectrum

Only the KNN method trained with the Correlation-based feature subset had a lower accuracy than the KNN method trained with all features. For the SVM all subsets increased accuracy. The results are further illustrated in Figure 24.

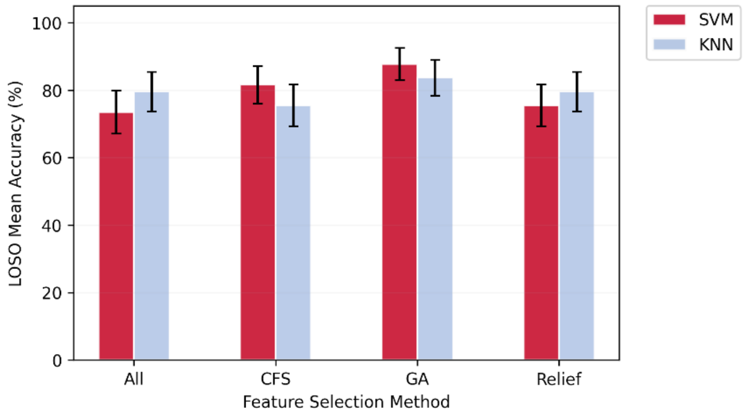

Figure 24. Mean *leave-one-subject-out* accuracies per feature selection method and classification model. Error bars represent the standard error of the mean. Feature Selection methods: **CFS** = Correlation-based feature selection, **GA** = Genetic Algorithm, **Relief** = Relief-based Algorithm. Models: **KNN** = k-nearest neighbor, **SVM** = Support Vector Machine. Note that for this feature selection and classification only relaxed trials have been included and stress has been labelled based on breathing variables. The state of the eyes has been neglected for this analysis. Hence, this dataset included 49 samples, 16 labelled as not stressed.

5. Discussion

The general goal of this thesis was to identify discriminative breathing and posture features that can predict stress. For that, several breathing and posture variables have been measured before and after relaxation inducing exercises. To verify the ground truth of the labelling the galvanic skin response (GSR) has been measured, but an ANOVA revealed no significant differences between the relaxed and stressed state. Only the factor eyes, open versus closed, had a significant effect on the sweat secretion of the participants. This may indicate that relaxation could not have been induced, but several other reasons are plausible. Firstly, the sample size has been relatively small, so differences between states may have not been detected due to insufficient statistical power. And secondly, GSR alone may not be indicative of stress. On one hand the sweat secretion can be easily confounded, for example through physical activity [4]. And on the other hand, the increase in skin conductance persists only 1 to 4 seconds [49]. As the recording of each state and eye condition lasted approximately 180 seconds and for this analysis the mean over the whole period has been used, the final score might have been relativized. The subjects in this study came into the room and were asked to undress. This might have caused an increase in sweat secretion, but only in the first few seconds. Participants may have still felt stressed and uncomfortable during the recording, but the initial GSR peak was perhaps already over. Future studies should consider more reliable measures to validate relaxation and stress, for example, a combination of heart rate, GSR, and acute stress questionnaires.

Assuming a successful induction of relaxation, this thesis separated the general goal in two parts: the identification of discriminative features and the predictive modelling. Four approaches have been used for feature selection, namely a Genetic Algorithm, a Correlation-based, and Relief-based approach, as well as a Decision Tree classifier which already embeds feature selection. In a first step the three self-implemented algorithms were applied to a benchmark dataset, to proof their functionality. The results showed that all three algorithms were able to decrease model complexity significantly by reducing the number of input variables, which maintained or often also increased model accuracy. Hence, all four feature selection methods have been applied to this thesis dataset. Breathing variables were found to be the most predictive, especially the mean of the tidal volume appeared in all selected feature subsets. However, as the relaxation intervention included deep breathing exercises, results may be biased. This is further discussed in the limitations. Stabilometry measures were also selected, but less often than breathing variables.

The binary prediction of stress versus no-stress, provided with the different feature subsets, achieved a *leave-one-subject-out* (LOSO) cross-validation accuracy of up to 85.42% with a Support Vector Machine (SVM). Considering only closed eyes trials, the LOSO accuracy was 85.42%, and for open eyes trials even up to 95.83%. These high values were both achieved with a SVM and can partially be explained since merging both, open and closed eyes, increases noise in stability measures. With open eyes visual input can be used for balance, whereas with closed eyes only proprioceptive information is available. This difference in information can cause differences in stabilometry measures, which are noise in terms of the classification as they occur in both classes. The feature subsets for both models were selected by the Genetic Algorithm and consisted of one breathing variable combined with several stabilometry measures. However, eliminating the breathing variable yielded an accuracy of 68.75%. Hence, it can be concluded that stabilometry measures alone are not sufficient for prediction.

Training a Decision Tree classifier with breathing and stabilometry measures achieved a mean LOSO accuracy of 78.12%. Among the features only one breathing variable, the mean tidal volume, and four stabilometry measure were present (see Figure 20). According to the tree a sample that has a low tidal volume is very likely from the stress condition, which makes sense as participants in the relaxed condition were asked to breathe deeply. Samples with higher values needed further investigation, for example, by splitting on the mean frequency of the spectrum in the medial-lateral displacement. A small frequency of sway in the x direction indicates a relaxed state. This is in line with findings from Tassani et al. (2019) who found that muscular tension, as it is the case in stressful situations, reduces stability of standing posture, which increases sway. The transversal displacement of the center-of-pressure (COP) is another stability splitting feature and further confirms this hypothesis, as samples with smaller values, and associated higher stability, are assigned to the relaxed state. The same applies for the transversal range, whereby subjects with smaller values are more likely to be assigned to the relaxed state.

Taken together, the accuracies found in this thesis are comparable to the ones found in literature and even outperformed some of them [4], [29], [39], [41]. The features that were identified as most predictive were breathing variables, which is in line with previous findings [27], [29], [41]. Nonetheless, the findings of this thesis indicate that adding stability measures contributes predictive power. Future work should therefore focus on how stability and posture measures could be integrated into stress detection, as adding them yields promising accuracy increases.

Nevertheless, several methodological limitations apply for this study. First of all, the assumed ground truth of stress versus no-stress. For this thesis it is assumed that participants were stressed since they were standing half-naked in the middle of a room watched by two unfamiliar laboratory assistants. But no real, controlled stressor occurred and the reaction to this new social interaction may be biased through inter-individual differences in social stress perception. One person is more sensitive to such a situation than the other. One option to consider these differences would be to include personality trait questionnaires, for example, the State-Trait Anxiety Inventory (STAI) [59]. This was initially planned for this study as well, but due to delays in data acquisition, it could not be used for this thesis. Furthermore, the stress during recording could not be verified. It was expected that sweat secretion before the relaxation induction is higher due to participants feeling uncomfortable, but this was not confirmed. To validate the labelling of stress versus no stress, future studies should apply self-reports before and after each standing period, and additional measures, such as the heart rate, should be used. This would increase validity of the classification labels and thus, would allow to draw more reliable and valid conclusions.

The main methodological limitation of this study, also regarding the ground truth of the labels, is the circular reasoning. Participants had to perform deep breathing and joint relaxation exercises prior to the relaxed recording, and where also asked to continue deep breathing during the recording. But then input features for prediction also included breathing variables. To clarify: Relaxation has been induced with breathing and then predicted with breathing variables. Hence, the prediction in this thesis was more about whether participants were able to successfully perform deep breathing, than to predict stress, as no further validation has been performed on perceived stress. That also explains why breathing variables were found to be the most discriminative. In this light the results of Figure 21 will also make sense, as of course the mean tidal, abdominal, and pulmonal ribcage volume increased in the second recording. Outliers in this graph may depict subjects that were unable to breathe deeply in the relaxed or performed deep breathing in the normal state. Nonetheless, stabilometry measures were also selected by the feature selection algorithms and therefore may contribute to prediction. The reason for them to be not sufficient in prediction could have been due to improper abdominal breathing of some participants.

This represents a further limitation regarding the ground truth of the labelling. Participants were asked to perform deep breathing and joint relaxation exercises, which aimed at relaxing them. However, in a study conducted by Tassani et al. (2019) 8 out of 30 subjects found it impossible or very difficult to perform abdominal breathing, and in a work of Ahmed et al. (2016), 9 out of 15 participants were unable to breathe deeply. In the relaxed state of this study 33 out of 49 subjects had a higher pulmonal ribcage volume than abdominal volume, even though they were asked to breathe abdominally. This indicates that most of the subjects were not able to perform deep breathing. For participants that have never done breathing exercises before, this task may therefore have even stressed them more, as it relaxed them. This hypothesis is further supported by the results of Khan et al. (2013) who concluded that inducing relaxation through deep breathing is not effective unless it is performed correctly. Moreover, the authors found that improper breathing even increased stress in 60% of the subjects. Since literature suggests that breathing abdominally induces relaxation [7], [19] and improper deep breathing even induces stress [60], an exploratory analysis has been performed in this study which labelled samples as stressed versus not stressed, based on whether a person was able to breathe deeply in the relaxed condition or not. Training a SVM with a subset of stability measures reached an accuracy of up to 87.76%. If breathing abdominally induces relaxation, and improper breathing is perceived as stressful, these results indicate that stress detection based on stability measures is possible, which further suggests a tight link between stability and relaxation/stress. Of course, further validation of perceived stress, for example, through questionnaires and physiological measures, is needed to draw valid conclusions.

A more common approach in stress detection studies is the induction of stress instead of relaxation induction. As described in the related works section several stress inducing paradigms. For a real-world stressor Healey & Picard (2005) proposed a driving protocol, where high stress is induced by letting participants drive through the city. The success of this intervention on perceived stress, as well as on physiology has been validated in several studies with questionnaires and physiological measures [24], [27], [41]. However, as the main goal of this study was to include stability measures, a real-world stressor may not be applicable. For the BYMBOS project a second session was planned applying the Trier Social Stress Test (TSST) [43], where stress is induced by asking participants to perform public speaking and mental arithmetic tasks. Due to delays in data acquisition this data was unfortunately not available. Nevertheless, there were no plans to record stability or breathing data during this session anyway, as public speaking may cause noise in recordings. Hence, to investigate the predictive power of stability measures in acute stress detection, the TSST method may also not be applicable. But besides the public speaking part of the TSST, there is the mental arithmetic task, which is also inducing physiological and psychological stress responses [1]–[3]. As this only requires thinking and a short period of answering, noise due to body movement may be reduced. Applying it only in the stress condition, may lead to distortions, as speaking also might change the respiratory pattern. For example, if a person answers on the arithmetic task, the change in the breathing pattern or stability might cause artifacts. These artifacts could help the algorithm classifying stress versus no stress but do not represent real-world variations as talking people might then always be classified as stressed. If arithmetic tasks are used to induce stress, these variations cannot be eliminated. However, they can be kept constant within subjects over the states by adding arithmetic tasks in both conditions, with the arithmetic tasks in the stressful condition inducing stress, whereas in the other condition they do not. Martin & Diverrez (2016) followed this approach and asked participants to answer arithmetic tasks in both conditions. In the situation without stress, the arithmetic task was presented, and the subjects heard a signal after approximately 3 seconds, after which they were supposed to answer. In contrast, in the stressful situation, they were supposed to answer before a signal sounded approximately one second after the calculation was presented. The effect of induced stress on the subjective feeling was validated with two questionnaires. Hence, this approach could be used in a future study with similar procedure and measurements as used in this study, to induce stress. Additionally, the relaxation induction could be enhanced by introducing a standardized approach, for example using audio guides. This would further reduce between subject differences. Combined with self-reports and physiological validation procedures this would better differentiate the stress and no stress condition. Furthermore, more participants are required in future studies to get more stable results. In particular, the imbalance of male and female subjects limits the generalizability of the results across gender. Hence, combining validated stress induction, standardized relaxation induction, and more samples with the study design of this thesis could give further and more validated insights into the predictive relationship between breathing, stability, and psychological stress.

Lastly, it should be mentioned that this study aimed at predicting acute psychological stress. However, one of the main goals of the BYMBOS project is to help students to acquire stress coping capabilities. Hence, a model that can predict stress resilience may be more appropriate from a practical point of view, as it would allow to identify markers indicating whether students are able to cope with stressful events or not. A candidate for such a marker would be the ability to breathe abdominally, as the results of this study further support the hypothesis that not everyone is able to perform deep breathing [18], [25], [60], which is known to be a stress regulating mechanism [7], [18]–[20]. Further biophysiological markers in resilience research include the cortisol reactivity and the heart rate variability [5], [38]. The effect of adding breathing and stability measure in predictive modelling remains to be investigated.

6. Conclusion

This exploratory analysis of the ongoing BYMBOS project aimed at identifying discriminative breathing and stability features that can predict stress. For that, several feature selection methods were applied. The results confirmed that these methods are an appropriate tool to reduce model complexity and can even increase prediction accuracy. Furthermore, it was found that stress prediction based on breathing and stability is possible, but stability measures alone were not sufficient. However, this might have been caused through an improper ground truth of stress, as many students were not able to breathe abdominally, and thus, relaxation induction might have failed. Changing the ground truth of stress to whether a person can perform deep breathing or not, increased predictive power of stability measures drastically. Moreover, findings of the Decision Tree classifier which mainly included stability measures indicate that relaxation is associated with higher stability. Taken together, these findings suggest a tight link between stability and psychological stress. Nevertheless, this study was part of an ongoing project and due to missing perceived stress validation, as well as due to the small sample size, results may not be representative. Hence, future studies should include more subjects and should consider a design which also induces stress, besides relaxation, to further confirm these hypotheses. Furthermore, more educational efforts should be made to teach students how to breathe deeply, as a high number of participating students was not able to breathe abdominally, which is a key factor in self-regulation. As this population is considerably exposed to stress, it is an important step towards better stress management, and thus, a higher quality of life.

List of Figures

List of Tables

References

[1] B. M. Kudielka, H. Hellhammer, and C. Kirschbaum, “Ten years of research with the trier social stress test.,” *Soc. Neurosci.*, pp. 56–83, 2007.

[2] J. Hellhammer and M. Schubert, “The physiological response to Trier Social Stress Test relates to subjective measures of stress during but not before or after the test,” *Psychoneuroendocrinology*, vol. 37, no. 1, pp. 119–124, 2012, doi: 10.1016/j.psyneuen.2011.05.012.

[3] A. P. Allen, P. J. Kennedy, J. F. Cryan, T. G. Dinan, and G. Clarke, “Biological and psychological markers of stress in humans: Focus on the Trier Social Stress Test,” *Neurosci. Biobehav. Rev.*, vol. 38, pp. 94–124, 2014, doi: 10.1016/j.neubiorev.2013.11.005.

[4] K. Plarre *et al.*, “Continuous inference of psychological stress from sensory measurements collected in the natural environment,” *Proc. 10th ACM/IEEE Int. Conf. Inf. Process. Sens. Networks, IPSN’11*, pp. 97–108, 2011.

[5] W. K. W. Lau, A. P. L. Tai, J. N. M. Chan, B. W. M. Lau, and X. Geng, “Integrative psycho-biophysiological markers in predicting psychological resilience,” *Psychoneuroendocrinology*, vol. 129, no. January, p. 105267, Jul. 2021, doi: 10.1016/j.psyneuen.2021.105267.

[6] T. G. Guilliams and L. Edwards, “Chronic stress and the HPA axis: Clinical assessment and therapeutic considerations,” *Stand.*, vol. 9, no. 2, pp. 1–12, 2010.

[7] V. Perciavalle *et al.*, “The role of deep breathing on stress,” *Neurol. Sci.*, vol. 38, no. 3, pp. 451–458, 2017, doi: 10.1007/s10072-016-2790-8.

[8] K. A. Muscatell and N. I. Eisenberger, “A Social Neuroscience Perspective on Stress and Health,” *Soc. Personal. Psychol. Compass*, vol. 6, no. 12, pp. 890–904, Dec. 2012, doi: 10.1111/j.1751-9004.2012.00467.x.

[9] M.-F. Marin *et al.*, “Chronic stress, cognitive functioning and mental health,” *Neurobiol. Learn. Mem.*, vol. 96, no. 4, pp. 583–595, Nov. 2011, doi: 10.1016/j.nlm.2011.02.016.

[10] T. Esch, G. B. Stefano, G. L. Fricchione, and H. Benson, “Stress in cardiovascular diseases,” *Med. Sci. Monit.*, vol. 8, no. 5, pp. 93–102, 2002.

[11] A. Hauke, J. Flintrop, E. Brun, and R. Rugulies, “The impact of work-related psychosocial stressors on the onset of musculoskeletal disorders in specific body regions: A review and metaanalysis of 54 longitudinal studies,” *Work Stress*, vol. 25, no. 3, pp. 243–256, 2011, doi: 10.1080/02678373.2011.614069.

[12] Y. Prins, L. Crous, and Q. Louw, “A systematic review of posture and psychosocial factors as contributors to upper quadrant musculoskeletal pain in children and adolescents,” *Physiother. Theory Pract.*, vol. 24, no. 4, pp. 221–242, 2008, doi: 10.1080/09593980701704089.

[13] K. Jacobs, S. Hudak, and J. McGiffert, “Computer-related posture and musculoskeletal discomfort in middle school students,” *Work*, vol. 32, no. 3, pp. 275–283, 2009, doi: 10.3233/WOR-2009-0826.

[14] K. Jacobs and N. A. Baker, “The association between children’s computer use and musculoskeletal

discomfort.," *Work*, vol. 18, no. 3, pp. 221–6, 2002, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/12441562.

[15] C. Regehr, D. Glancy, and A. Pitts, "Interventions to reduce stress in university students: A review and meta-analysis," *J. Affect. Disord.*, vol. 148, no. 1, pp. 1–11, May 2013, doi: 10.1016/j.jad.2012.11.026.

[16] M. C. Pascoe, S. E. Hetrick, and A. G. Parker, "The impact of stress on students in secondary school and higher education," *Int. J. Adolesc. Youth*, vol. 25, no. 1, pp. 104–112, Dec. 2020, doi: 10.1080/02673843.2019.1596823.

[17] L. Zou *et al.*, "Effects of Mind–Body Exercises (Tai Chi/Yoga) on Heart Rate Variability Parameters and Perceived Stress: A Systematic Review with Meta-Analysis of Randomized Controlled Trials," *J. Clin. Med.*, vol. 7, no. 11, p. 404, Oct. 2018, doi: 10.3390/jcm7110404.

[18] S. Tassani, J. M. Font-Llagunes, M. Á. González Ballester, and J. Noailly, "Muscular tension significantly affects stability in standing posture," *Gait Posture*, vol. 68, no. November 2018, pp. 220–226, 2019, doi: 10.1016/j.gaitpost.2018.11.034.

[19] X. Ma *et al.*, "The Effect of Diaphragmatic Breathing on Attention, Negative Affect and Stress in Healthy Adults," *Front. Psychol.*, vol. 8, no. JUN, pp. 1–12, Jun. 2017, doi: 10.3389/fpsyg.2017.00874.

[20] S. I. Hopper, S. L. Murray, L. R. Ferrara, and J. K. Singleton, "Effectiveness of diaphragmatic breathing for reducing physiological and psychological stress in adults," *JBI Database Syst. Rev. Implement. Reports*, vol. 17, no. 9, pp. 1855–1876, Sep. 2019, doi: 10.11124/JBISRIR-2017-003848.

[21] S. Tassani, "BYMBOS - Breathing Dynamic Modelling for Body Mind Interaction in Students," 2020. https://www.upf.edu/web/bymbos.

[22] Y. Shi *et al.*, "Personalized Stress Detection from Physiological Measurements," *Second Int. Symp. Qual. Life Technol.*, 2010, [Online]. Available: http://www.shrs.pitt.edu/qolt/qolt.aspx?id=2212.

[23] K. Soman, V. Alex, and C. Srinivas, "Analysis of physiological signals in response to stress using ECG and respiratory signals of automobile drivers," in *2013 International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s)*, Mar. 2013, pp. 574–579, doi: 10.1109/iMac4s.2013.6526476.

[24] J. A. Healey and R. W. Picard, "Detecting Stress During Real-World Driving Tasks Using Physiological Sensors," *IEEE Trans. Intell. Transp. Syst.*, vol. 6, no. 2, pp. 156–166, Jun. 2005, doi: 10.1109/TITS.2005.848368.

[25] B. Ahmed, H. M. Khan, J. Choi, and R. Gutierrez-Osuna, "ReBreathe: A Calibration Protocol that Improves Stress/Relax Classification by Relabeling Deep Breathing Relaxation Exercises," *IEEE Trans. Affect. Comput.*, vol. 7, no. 2, pp. 150–161, Apr. 2016, doi: 10.1109/TAFFC.2015.2459682.

[26] S. S. Panicker and P. Gayathri, "A survey of machine learning techniques in physiology based mental stress detection systems," *Biocybern. Biomed. Eng.*, vol. 39, no. 2, pp. 444–469, 2019, doi:

10.1016/j.bbe.2019.01.004.

[27] A. Ghaderi, J. Frounchi, and A. Farnam, “Machine learning-based signal processing using physiological signals for stress detection,” *2015 22nd Iran. Conf. Biomed. Eng. ICBME 2015*, no. November 2015, pp. 93–98, 2016, doi: 10.1109/ICBME.2015.7404123.

[28] P. Bobade and M. Vani, “Stress Detection with Machine Learning and Deep Learning using Multimodal Physiological Data,” *Proc. 2nd Int. Conf. Inven. Res. Comput. Appl. ICIRCA 2020*, pp. 51–57, 2020, doi: 10.1109/ICIRCA48905.2020.9183244.

[29] M. Abouelenien, M. Burzo, and R. Mihalcea, “Human acute stress detection via integration of physiological signals and thermal imaging,” *ACM Int. Conf. Proceeding Ser.*, vol. 29-June-20, 2016, doi: 10.1145/2910674.2910705.

[30] B. J. Park, E. H. Jang, S. H. Kim, C. Huh, and J. H. Sohn, “Feature selection on multi-physiological signals for emotion recognition,” *Proc. - 2nd Int. Conf. Eng. Ind. ICEI 2011*, pp. 71–76, 2011.

[31] P. Gong, H. T. Ma, and Y. Wang, “Emotion recognition based on the multiple physiological signals,” in *2016 IEEE International Conference on Real-time Computing and Robotics (RCAR)*, Jun. 2016, vol. 44, no. 4, pp. 140–143, doi: 10.1109/RCAR.2016.7784015.

[32] J. Cai, J. Luo, S. Wang, and S. Yang, “Feature selection in machine learning: A new perspective,” *Neurocomputing*, vol. 300, pp. 70–79, 2018, doi: 10.1016/j.neucom.2017.11.077.

[33] R. J. Urbanowicz, M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore, “Relief-based feature selection: Introduction and review,” *J. Biomed. Inform.*, vol. 85, no. July, pp. 189–203, 2018, doi: 10.1016/j.jbi.2018.07.014.

[34] F. Tan, X. Fu, Y. Zhang, and A. G. Bourgeois, “A genetic algorithm-based method for feature subset selection,” *Soft Comput.*, vol. 12, no. 2, pp. 111–120, Sep. 2007, doi: 10.1007/s00500-007-0193-8.

[35] A. Bommert, X. Sun, B. Bischl, J. Rahnenführer, and M. Lang, “Benchmark for filter methods for feature selection in high-dimensional classification data,” *Comput. Stat. Data Anal.*, vol. 143, p. 106839, Mar. 2020, doi: 10.1016/j.csda.2019.106839.

[36] M. Hall, “Correlation-based feature selection of discrete and numeric class machine learning,” in *Proceedings of the 17th International Converence on Machine Learning*, 2000, pp. 359–366.

[37] O. Kramer, *Genetic Algorithm Essentials*, vol. 679. Cham: Springer International Publishing, 2017.

[38] K. Schultebraucks and I. R. Galatzer-Levy, “Machine Learning for Prediction of Posttraumatic Stress and Resilience Following Trauma: An Overview of Basic Concepts and Recent Advances,” *J. Trauma. Stress*, vol. 32, no. 2, pp. 215–225, Apr. 2019, doi: 10.1002/jts.22384.

[39] N. Martin and J.-M. Diverrez, “From Physiological Measures to an Automatic Recognition System of Stress,” in *Communications in Computer and Information Science*, C. Stephanidis, Ed. Springer, 2016, pp. 172–176.

[40] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, “The WEKA Data Mining Software: An Update,” *SIGKDD Explor.*, vol. 11, no. 1, 2009.

[41] P. Napoletano and S. Rossi, “Combining heart and breathing rate for car driver stress recognition,” *IEEE Int. Conf. Consum. Electron. - Berlin, ICCE-Berlin*, vol. 2018-Septe, 2018, doi: 10.1109/ICCE-Berlin.2018.8576164.

[42] A. Riera *et al.*, “Electro-Physiological Data Fusion for Stress Detection,” in *Annual Review of Cybertherapy and Telemedicine*, B. K. Wiederhold and G. Riva, Eds. 2012 Interactive Media Institute, 2012, pp. 228–232.

[43] C. Kirschbaum, K.-M. Pirke, and D. H. Hellhammer, “The ‘Trier Social Stress Test’ – A Tool for Investigating Psychobiological Stress Responses in a Laboratory Setting,” *Neuropsychobiology*, vol. 28, no. 1–2, pp. 76–81, 1993, doi: 10.1159/000119004.

[44] P. Schmidt, A. Reiss, R. Duerichen, C. Marberger, and K. Van Laerhoven, “Introducing WESAD, a Multimodal Dataset for Wearable Stress and Affect Detection,” in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, Oct. 2018, pp. 400–408, doi: 10.1145/3242969.3242985.

[45] T. E. Prieto, J. B. Myklebust, R. G. Hoffmann, E. G. Lovett, and B. M. Myklebust, “Measures of postural steadiness: Differences between healthy young and elderly adults,” *IEEE Trans. Biomed. Eng.*, vol. 43, no. 9, pp. 956–966, 1996, doi: 10.1109/10.532130.

[46] A. Aliverti and A. Pedotti, “Opto-electronic plethysmograpy,” *Monaldi Archives for Chest Disease - Pulmonary Series*, vol. 59, no. 1. Springer Milan, Milano, pp. 12–16, 2003, doi: 10.1007/978-88-470-2916-3_5.

[47] S. J. Cala *et al.*, “Chest wall and lung volume estimation by optical reflectance motion analysis,” *J. Appl. Physiol.*, vol. 81, no. 6, pp. 2680–2689, Dec. 1996, doi: 10.1152/jappl.1996.81.6.2680.

[48] A. Aliverti, R. Dellacà, P. Pelosi, D. Chiumello, L. Gattinoni, and A. Pedotti, “Compartmental analysis of breathing in the supine and prone positions by optoelectronic plethysmography,” *Ann. Biomed. Eng.*, vol. 29, no. 1, pp. 60–70, 2001, doi: 10.1114/1.1332084.

[49] D. Caruelle, A. Gustafsson, P. Shams, and L. Lervik-Olsen, “The use of electrodermal activity (EDA) measurement to understand consumer emotions – A literature review and a call for action,” *J. Bus. Res.*, vol. 104, no. July, pp. 146–160, Nov. 2019, doi: 10.1016/j.jbusres.2019.06.041.

[50] A. Greco, G. Valenza, and E. P. Scilingo, *Advances in Electrodermal Activity Processing with Applications for Mental Health*. Cham: Springer International Publishing, 2016.

[51] C. M. Bishop, *Pattern Recognition and Machine Learning*. 2006.

[52] J. C. C. Burges, “A Tutorial on Support Vector Machines for Pattern Recognition,” *Data Min. Knowl. Discov.*, vol. 2, pp. 121–167, 1998.

[53] F. Pedregosa, G. Varoquaux, L. Buitinck, G. Louppe, O. Grisel, and A. Mueller, “Scikit-learn: Machine Learning in Python,” *J. Mach. Learn. Res.*, vol. 12, no. 1, pp. 2825–2830, 2011, Accessed: Jun. 12, 2021. [Online]. Available: http://scikit-learn.sourceforge.net.

[54] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning*. Cambridge: Cambridge University Press, 2014.

[55] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, “An introduction to decision tree modeling,” *J. Chemom.*, vol. 18, no. 6, pp. 275–285, Jun. 2004, doi: 10.1002/cem.873.

[56] K. Kira and L. A. Rendell, “A Practical Approach to Feature Selection,” in *Machine Learning Proceedings 1992*, Elsevier, 1992, pp. 249–256.

[57] I. Kononenko, “Estimating attributes: Analysis and extensions of RELIEF,” in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 784 LNCS, 1994, pp. 171–182.

[58] S. Sayed, M. Nassef, A. Badr, and I. Farag, “A Nested Genetic Algorithm for feature selection in high-dimensional cancer Microarray datasets,” *Expert Syst. Appl.*, vol. 121, pp. 233–243, May 2019, doi: 10.1016/j.eswa.2018.12.022.

[59] C. D. Spielberger, “State-Trait Anxiety Inventory for Adults (STAI-AD),” *APA PsycTests*, 1983.

[60] H. M. Khan, B. Ahmed, Jongyoon Choi, and R. Gutierrez-Osuna, “Using an ambulatory stress monitoring device to identify relaxation due to untrained deep breathing,” in *2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Jul. 2013, pp. 1744–1747, doi: 10.1109/EMBC.2013.6609857.