Detection of Teeth Grinding and Clenching using Surface Electromyography

Hella Toto-Kiesa

Master's thesis

Detection of Teeth Grinding and Clenching using Surface Electromyography

Hella Toto-Kiesa

July 29th, 2020

Chair of Data Processing
Technische Universität München

Hella Toto-Kiesa. *Detection of Teeth Grinding and Clenching using Surface Electromyography.* Master's thesis, Technische Universität München, Munich, Germany, 2020.

Supervised by Prof. Dr.-Ing. Klaus Diepold and Matthias Kissel, M.Sc. and Michael Moosmeier, M.Sc.; submitted on July 29th, 2020 to the Department of Electrical and Computer Engineering of the Technische Universität München.

Chair of Data Processing, Technische Universität München, 80290 München, Germany, http://www.ldv.ei.tum.de/.

Eidesstattliche Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Mir ist bekannt, dass nur diese digitale Version gültig ist und bewertet wird. Weitere digitale Versionen oder gedruckte Versionen werden nicht für die Bewertung berücksichtigt.

München, 29.07.2020 H. Toto-Kiesa

(Ort, Datum) (Hella Toto-Kiesa)

Abstract

This thesis presents the detection and analysis of Teeth Grinding and Clenching by using Surface Electromyography (sEMG). The question of the existence of a relationship between data and conscious teeth grinding and clenching is discussed in this thesis, and will be answered in relation to the relevant research questions. Previous work has shown that sEMG along with machine learning techniques are suitable for the detection of medical disorders e. g. gait disorders, for the development of electronic wheelchairs or even prostheses. Most techniques for the detection of bruxism, which is defined as extreme teeth grinding and clenching, focus on the determination of threshold values for automatic detection. In this thesis the focus is set on finding appropriate machine learning algorithms for the classification of EMG signals acquired from the temporalis muscle. It is classified whether teeth are being grinded and clenched or not. The potential of Logistic Regression, Support Vector Machine and Random Forest classifiers along with different sets of features is evaluated. Additionally, it is examined whether classifier calibration can improve the model in terms of generalizability. After evaluation of different techniques it is empirically shown that the random forest model with a feature set of eight time-domain features works best. It is also shown, that calibration with Isotonic Regression is better suitable for the detection of teeth grinding and clenching than calibration using Platt Scaling. However, the calibration curves presented later in this thesis show that both techniques are not optimal.

Contents

Acronyms

BS Brier Score

DBSCAN Density-based spatial clustering of applications with noise

DWT Discrete Wavelet Transform

EMG Electromyography

IEMG Integrated EMG

LOG Log-Detector

MUAP Motor Unit Action Potentials

PCA Principal Component Analysis

RBF Radial Basis Function

RMS Root Mean Square

sEMG Surface Electromyography

SSC Slope Sign Change

SSI Simple Square Integral

SVM Support Vector Machine

TD Time Domain

TFD Time Frequency Domain

Var Variance of EMG

WAMP Willson Amplitude

WL Waveform Length

1. Introduction

According to the study conducted by Manfredini et al. in 2013 around $8 - 31\,\%$ of the worldwide population suffers from the oral parafunctional activity called bruxism [10]. Bruxism can occur during day and night times i. e. while the person is awake or asleep. The parasomnia *sleep bruxism* that is characterized by nocturnal teeth clenching and grinding is not categorized as a disease. However, it can either be an indicator for a psychological disorder or have physical impacts on the affected person. Possible risk factors are craniomandibular disorders (e.g. headache), premature loss of teeth, dental damages and morning soreness [15].

1.1. Motivation

Countermeasures can be taken quickly if diagnosis is made promptly, in order to reduce the risk of consequential health damage. Diagnosis can be achieved by monitoring and/or analyzing the activity of the masticatory muscles. According to the guideline for the diagnosis and treatment of bruxism by the *German Corporation for Functional Diagnosis and Therapy* (DGFDT) [17] six diagnosis techniques are commonly applied in science and practice.

The first technique is the anamnestic questionnaire. An **anamnestic questionnaire** can be used to asses whether a patient clenches or grinds consciously or unconsciously with their teeth. As this method relies on the patient's or their relative's perception, it is not accurate and tends to overdiagnosis. It should therefore be applied primarily for screening.

Clinical examination, the second diagnosis technique, is similarly to anamnesis recommended only for screening, because medical symptoms are not reliable for accurate diagnosis.

Moreover, there exist particular **splints** which gradually change color depending on the depth of the abrasion due to teeth grinding. The technique has a sensitivity of 79,2 % and specificity of 95 % for 2900 pixelscores (which are score from abrasion area and number of layers). However, it can promote neuromuscular changes related to the changed occlusal height. Moreover, it is important to mention that this technique does not record jaw clenching.

The third technique which is known as **Polysomnography (PSG)** is the standard diagnosis technique for sleep bruxism. It is conducted in a sleep laboratory, where physiological parameters (e. g. Electroencephalography (EEG), Electromyography (EMG), Electrocardiography (ECG), etc.) are registered. Video or audio may be additionally recorded. Although it is highly popular due to its precision, it is generally used only on a small number of participants during studies, because of the high technical, financial and time expense. Furthermore, the test person could show different habits during the study when compared to the habits demonstrated in their usual sleep environment.

A good alternative are **wearable EMG devices** which capture the muscle activity using the potential difference between muscle and skin. The devices provide good results and are less expensive than PSG. Moreover, they are not restricted to a particular location.

Biofeedback is another technique that has been qualified as promising but has not yet been sufficiently evaluated. It consists of a sensor (usually EMG) and an actuator (e. g. acoustic signal) for feedback, which makes it not only applicable for diagnosis but also for therapy. A biofeedback system can be considered an extension of wearable EMG devices. The aim of this thesis is to develop a concept for a EMG-Biofeedback device for detection of teeth grinding and clenching. The question whether Biofeedback is a robust technique is not discussed.

1.2. Literature and Patent Review

Classification based on surface electromyography (sEMG) has been a popular research topic in the past years. It has found, inter alia, its application in the detection of gait disorders [3], facial gesture recognition for stress monitoring [16], upper limb prostheses [20] but also in dentistry, e. g. for the detection of sleep bruxism [13].

Castroflorio et al. [2] used masseter surface EMGs with bipolar concentric electrodes and ECG with monopolar electrodes on the clavicular regions to detect sleep bruxism episodes in the natural sleep environment (at the patient's home). If the EMG signal exceeds the threshold by $10\,\%$ and the heart rate increases by more than $25\,\%$ $1\,s$ before the rise of the EMG amplitude, a bruxism episode is detected. An additional neural network was used to classify the subjects as bruxers and non-bruxers. The classification error between both classes was $1\,\%$.

There has also been some patent applications for the detection of bruxism using sEMG. Under the application number CN1810319A an electromechanical bruxism monitoring and treating equipment is introduced. Detection occured if EMG exceeds $20\,\mu V$ and a period longer than one second [26]. A similar tool was invented by Stephen H. Ober [14]. The biofeedback apparatus detected bruxism if an adjustable threshold was exceeded. The number of detection can be stored and stimulus on the temporo-mandibular joint is applied to open the jaw (biofeedback). In 2007, Momen et al. [11] applied for a patent for an apparatus that classifies signals from user-selected intentional movements in real-time. Four different types of classifiers have been proposed, namely *linear , Artificial neural networks (ANN), k-Nearest Neighbors (kNN) and Fuzzy classifiers*. Different feature sets were examined here: time-domain features and various wavelet transform techniques. In 2009, the Thumedi & Co KG GmbH applied for a patent for "measuring, analysis and biofeedback device". The device contains at least one signal amplifier, electrodes and a signal evaluation unit. The evaluation is software-based and takes into account physiological effects [24].

1.3. Research Questions

As presented in Section 1.2, muscular activity can be measured with the help of electromyography. Using suitable machine learning algorithms, I endeavour to find a relation between the measured data and unconscious teeth grinding and clenching. The following research questions are covered in this thesis:

1. Is there a the dependency between the jaw muscle activity and the EMG sensor data?
2. Which Machine Learning methods are best suitable for the detection of teeth grinding and clenching?
3. How does calibration influence the models generalizability?

1.4. Approach

Data is acquired with the help of a 3-channel EMG sensor (see Section 2.1). It is processed and analysed according to the strategy presented in the section on Experimental Design. The processed data is used to train the classifiers described in Section 2.5. Their performances are compared and evaluated in Chapter 4. The results are then used to develop a concept for an EMG-Biofeedback device, as presented in Section 4.4. The final outcome is critically assessed in the Discussion chapter.

2. Technical Background

A biofeedback system consists of several elements for the acquisition and processing of electrical signals recorded from the human body (biosignals). Generally, a sensor (measurement device) is needed to retrieve the biosignal, which is then processed such that the data is suitable for the controller unit. The controller is then charged to control the actuator, which thereafter sends the feedback to the human. Figure 2.1 shows a general block diagram of the components of a biofeedback system. This chapter provides a theoretical overview of applied methodologies.

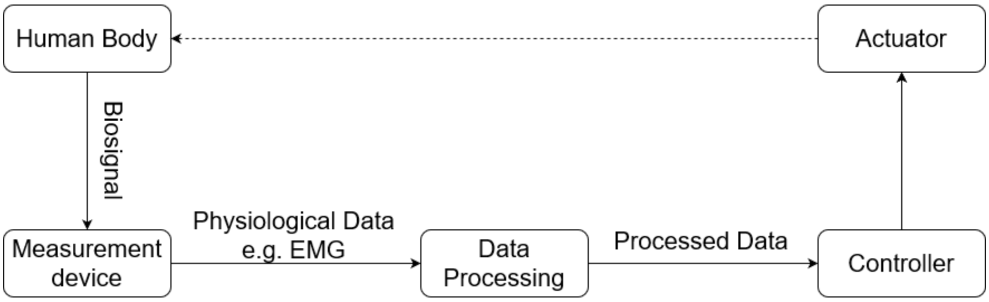


Figure 2.1: General block diagram of a biofeedback system.

2.1. Electromyography

The term electromyography (EMG) is composed by the words electro, myo (*mys* greek for *muscle*) and -graphy (*grapho* greek for *something written*). It describes the recording of the electrical activity of a muscle. The movement of a muscle or a muscle group is controlled by "motor neuron signals originating from the central nervous system (CNS)" [22]. EMG signals contain anatomical and physiological information of the muscle. The data can be acquired using EMG devices and can for example be used to develop a Myoelectric Control System (MCS) for example, a system that uses EMG to control electrical devices (e. g. electrical wheelchairs) [3]. A distinction is made between invasive and non-invasive EMG. The former records the electrical activity by inserting a needle electrode into the respective muscle. Surface Electromyography is a non-invasive method, where adhesive button electrodes are placed on the skin above the respective muscle. This section provides an overview on the required elements for EMG.

2.1.1. EMG Signal Generation

Due to the ionic difference between inner and outer space of a muscle cell, there exists a resting potential of approximately $-80\ to\ -90\,mV$ at the muscle fiber membrane, when it is not contracted. This difference is sustained by the ion pump resulting in a negative intra-cellular charge when compared to the skin. During contraction Na^+ ions flux through the semi-permeable membrane which changes the **action potential** from $-80\,mV$ to $+30\,mV$ (**Depolarization**). The potential is restored by a backward exchange within the ion pump, i. e. increased influx of K^+ ions (**Repolarization**). Figure 2.2 shows an illustration of the depolarization and repolarization cycle. The depolarization zone is approximately $1-3\,mm^2$ and travels along the muscle fiber with a velocity of $2-6\,m/s$. For bipolar electrodes, the potential difference during the movement of the depolarization zones is measured between the electrodes. A motor unit, which is defined as "the smallest function unit to describe the neural control of the muscular contraction process", may contain many skeletal muscle fibers. Therefore, the overall action potentials are super-positioned to the so called Motor Unit Action Potentials (MUAP) [8]. Equation 2.1 shows a simple EMG signal model, where $EMG(n)$ is the sampled EMG signal, α_r the MUAP, $x(n)$ the firing impulse, w_n white Gaussian noise and N the number of motor unit firing at a particular time [22].

$$EMG(n) = \sum_{i=1}^{N-1} \alpha_i(r)x_i(n-r) + w_n \tag{2.1}$$

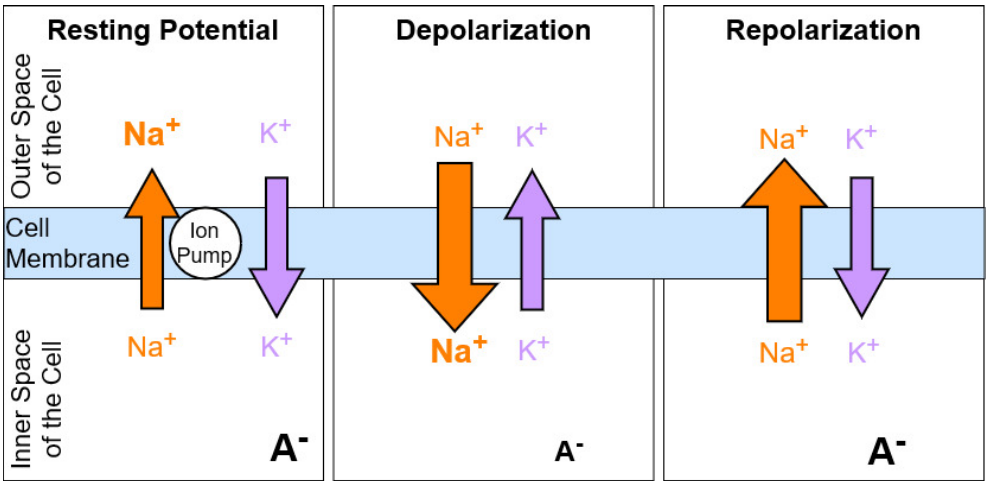


Figure 2.2: Illustration of the depolarization and repolarization cycle of the cell membrane, according to: [22] Fig. 6, page 7

2.1.2. Noise Sources and Reduction

For an accurate biofeedback system reliable EMG data is required. Thus, noise sources in EMG signals need to be considered and possible countermeasures need to be applied [3].

- Electronic devices produce **internal noises** ranging from 0 to several thousands of Hz. For the experimental design described in section 3.3, it should be kept in mind, that electrodes made of silver chloride generally provide an adequate Signal-to-Noise Ratio (SNR).
- **Movement artifacts** are another source of noises. They range from $1 - 10\,Hz$ and can either originate from the movement sensor cables or the muscle itself. When activating the muscle its length decreases which leads to the movement of muscle, skin and electrodes. The use of a conductive gel layer between skin and electrode is recommended for the reduction of these artifacts.
- The human body has the same properties as an antenna. It is surrounded by electric and magnetic radiation causing **electromagnetic noise**. The noise can be reduced using a high pass filter.
- **Cross Talk** (CT) is caused by undesired EMG signals from other muscle groups. Using a bony area as reference point and the use of small electrodes can minimize this.
- Moreover, it is recommended to filter frequencies ranging from $0-20\,Hz$ as they are unstable. This is primarily due to the quasi-randomness of the firing rate of the motor units, to whom they are related.
- **Electrocardiographic (ECG) artifacts** can also induce noise, because the electrical activity of the heart interfers with the surface EMG. The impact depends on the relative position of the corresponding muscle to the heart.

2.2. Feature Extraction

EMG signals do not usually contain interpretable information individually. Therefore, it is recommended to work with their representation vectors seen as a whole. This is done by extracting features from them, which will later be fed to the classifier. The goal is to find a feature set that maximizes the ability to seperate the data into the corresponding classes. Feature extraction is possible in different domains. Given the fact that myoelectric signals are time functions, they can be described in terms of amplitude, frequency or phase. Consequently, time, frequency and time-frequency domain features are possible [25]. As frequency domain features do not contain time information and are preferably used for spectral analysis needed to investigate muscle fatigue for example, this work will focus on time and time-frequency domain features.

2.2.1. Time Domain Features

The processing of the data in the time domain (TD) is useful as signals are usually already sampled in the time domain. This voids the need of a conversion and is thus simpler to extract. Moreover, it is less prone to noise. Nonetheless, myoelectric signal have non-stationary properties, which lead to time-dependent and varying statistical properties.

There exists a long list of possible TD features, but only few are compatible for the classification of EMG signals. Features based on the following statistical values are not suitable as they provided poor results in the study conducted by Phinyomark et al. [18], as stated in [3].

- Mean Frequency (MNF)
- Median Frequency (MDF)
- Mean Peak Frequency (MPF)
- Mean Power (MNP)
- Time-to-peak Force (TTP)
- Spectral Moments
- Frequency Ratio (FR)
- Power Spectrum Ratio (PSR) and
- Variance of Central Frequency (VCF)

A selection of the most common TD features [25], that where used in this thesis, and their representation is shown below.

- Integrated EMG (IEMG)
  IEMG is defined as the area under the curve of the rectified EMG signal. It is calculated as the integral of the absolute value of the signal. Its mathematical representation is shown in Equation 2.2. Where N is the signal length and x_n an EMG signal segment.
  $$IEMG = \sum_{n=1}^{N} |x_n| \tag{2.2}$$
- Simple Square Integral (SSI)
  SSI is the sum of the squared values of the EMG signal amplitude. It corresponds to the signal's energy. Equation 2.3 shows its mathematical expression.
  $$SSI = \sum_{n=1}^{N} |x_n|^2 \tag{2.3}$$

- Variance of EMG (Var)
  Variance is defined as the mean of the squared values of the deviation from the variable's mean. It corresponds to power and can be expressed as

$$VAR = \frac{1}{N-1}\sum_{n=1}^{N} x_n^2 \tag{2.4}$$

- Root Mean Square Value (RMS)
  RMS is calculated as the square root of the mean of the squared values (see eq. 2.5). It represents the constant force and contraction without fatigue.

$$RMS = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N} x_n^2} \tag{2.5}$$

- Waveform Length (WL)
  WL is the accumulated waveform length above a time segment and represents the complexity of the waveform in each segment. It is expressed as

$$WL = \sum_{n=1}^{N} |x_n - x_{n-1}| \tag{2.6}$$

- Log-Detector (LOG)
  The LOG feature corresponds to an estimate of the muscle contraction strength. It is defined as

$$LOG = \exp^{\frac{1}{N}\sum_{n=1}^{N} log|x_n|} \tag{2.7}$$

- Willson Amplitude (WAMP)
  WAMP corresponds to the number of times the amplitude of the absolute difference between two adjacent samples is higher than a threshold L, which is linked to the triggering of the action potentials of the motor unit and the muscle contraction force. It can be interpreted as the EMG signal frequency and is defined as

$$WAMP = \sum_{n=1}^{N-1} f\,|x_{n+1} - x_n| \ , \tag{2.8}$$

$$\text{where } f(x) = \begin{cases} \\ 1, & \text{if x} \geq \text{threshold,} \\ 0, & \text{otherwise} \end{cases}$$

- Slope Sign Change (SSC)
  Similar to WAMP, SSC provides a description of the EMG signal frequency. It

is defined as the number of times that a slope sign change occurs. This is calculated using three consecutive samples x_{k-1}, x_k, x_{k+1} as depicted in Equation 2.9.

$$SSC = \sum_{n=2}^{N-1} f\,|(x_n - x_{n-1}) \times (x_n - x_{n+1})|$$
$$f(x) = \begin{cases} \\ 1, & \text{if x} \geq \text{threshold,} \\ 0, & \text{otherwise} \end{cases} \tag{2.9}$$

2.2.2. Time Frequency Domain Features

In [12] Time Frequency Domain (TFD) features are described as a great alternative to the aforementioned features with the following limits. First, the frequency domain does not provide any time information, which is needed for classification, and second the time domain characteristics are comparatively weak. Although TFD features are computationally more costly than TD features, they can still meet real-time requirements. The problem with TFD features is, that they are highly dimensional which results in a high resolution of vectors.

A popular example of a TFD feature which is also used in this work, is the Wavelet Transform (WT). A distinction is made between discrete and concrete wavelet transform (DWT and CWT respectively) [3]. DWT is well suited for non-stationary signals like sEMG. It has a low processing time and yields high-dimensional feature vectors. The computation is done by successive low-pass and high-pass filtering in the discrete-time domain (see eq. 2.10).

$$x(t) = \sum_{k=-\infty}^{\infty} \sum_{t=-\infty}^{\infty} d(k,l) 2^{\frac{k}{2}} \psi(2^{-k}t - 1) \tag{2.10}$$

where $a = 2k$, $b = 2kl$ and $d(k,l)$ is a sampling of $W(a,b)$ at discrete points k and l.

However, CWT is more consistent and less time consuming compared to DWT, because the down-sampling step is skipped. Equation 2.11 shows the general expression of CWT, where a is the scale, b the time-location and $\psi(t)$ the "mother wavelet" which can be interpreted as a band-pass function.

$$\psi(a,b) = \frac{1}{\sqrt{|a|}} \psi(\frac{t-b}{a}) \tag{2.11}$$

2.3. Dimensionality Reduction with the Principal Component Analysis

Principal Component Analysis (PCA) is a method for dimensionality reduction. It finds its application in multivariate classification problems with a high number of correlated variables. The aim is to reduce the dimensionality of the data set without losing too much information. The resulting sets are called *principal components*. The algorithm is as follows [1].

Let X be a multivariate data matrix with n rows and p columns. One data row represents different measurements on a subject.

- Standardize matrix $\mathbf{X}$ such that the mean per column is 0 and the variance is 1. Call this matrix $\mathbf{Z}$. PCA aims to derive a linear function y for each vector variable $\mathbf{z_i}$ (column of $\mathbf{Z}$) with maximized variance.
- The linear function $\mathbf{y}$ is a component of $\mathbf{z}$. The computation of a single element for the j^{th} $\mathbf{y}$ vector is done by $\mathbf{y} = \mathbf{z}\mathbf{v}'$ where $\mathbf{v}'$ is a column vector of $\mathbf{V}$ and $\mathbf{V}$ the $p \times p$ eigenvector matrix. $\mathbf{z}$ and $\mathbf{v}'$ are of dimension $1 \times p$ and $p \times 1$ respectively. The i^{th} element of $\mathbf{y_j}$, $j = 1, \ldots, p$ is:

  $$y_{ij} = v'_1 z_{1i} + v'_2 t_{2i} + \cdots + v'_p z_{pi}$$

  Which becomes

  $$\mathbf{Y} = \mathbf{Z}\mathbf{V}$$

  for all of the y.
- Because the mean of each column is 0, the mean $m()$ of $\mathbf{y}$ is $\mathbf{m}(\mathbf{y}) = \mathbf{V}'\mathbf{m}(\mathbf{z}) = 0$
- The dispersion matrix of y is

  $$\mathbf{D_y} = \mathbf{V}'\mathbf{D_z}\mathbf{V} = \mathbf{V}'\mathbf{R}\mathbf{V}$$

- The dispersion matrix D_z of a standardized variable is a correlation matrix. Hence R is the correlation matrix of $\mathbf{z}$.

2.4. Clustering with DBSCAN

Clustering algorithms group data according to their similarity or membership, which is defined by different criteria, such as the distances between cluster centers or points.

[1] https://www.itl.nist.gov/div898/handbook/pmc/section5/pmc55.htm

The *Density-based spatial clustering of applications with noise* (DBSCAN) is based on the density of the input, i. e. clusters are formed in regions with high density of data within lower-density regions. The algorithm requires two input parameters. The minimum number of samples required for a point to be considered a core point ($min\_samples$) and the maximum distance between two samples (ϵ).

Every sample $\vec{x}_n$ can be categorized as one of the three types:

- **Core point** A core point has at least $min\_sample - 1$ other points within a circle of radius ϵ.
- **Border point** A border point is a point, that is not a core point and has at least one core point within the distance of ϵ.
- **Noise point** A point that is neither a border nor a core point is called noise point.

The DBSCAN algorithm is as follows [5]:

1. Label each point as one of the types presented above.
2. Eliminate the noise points
3. Connect all core points that are within the distance of ϵ from each other with edges.
4. Each connected component becomes one cluster.
5. Border points are assigned to a cluster that is associated with one of its nearby core points.

2.5. Classification Techniques

After pre processing, the data is used to train a classifier, which will be able to predict whether a test subject clenches or grinds with their teeth. There are several classification techniques used in the literature for the classification of EMG signals. Their performances usually depend on the intended application. This thesis will focus on the following techniques, which have proven to work well in similar applications. Support Vector Machines (SVM) for stress monitoring [16], Logistic Regression for the classification of diseases [7] and Random Forest for the classification of upper limb motions [19].

2.5.1. Logistic Regression

Logistic regression is a linear classification technique that models the probabilities of specific outcomes using the logistic function $f(z) = \frac{1}{1+e^{-z}}$. With z being the linear sum of the parameter α and the products of the parameters β_i and the independent variables X_i, we obtain the following logistic model:

$$f(z) = P(Y = 1|X_1, X_2, ..., X_k) = \frac{1}{1 + e^{-(\alpha+\sum\beta_i X_i)}} \tag{2.12}$$

The unknown parameters α and β_i are estimated such that the classification error is minimized. The class of the data is predicted using the resulting probability. This is done by comparing it to a preset threshold [7].

To prevent the model from overfitting we can apply regularization, which solves the following optimization problems [2]. Where w are the weights, C the regularization parameter, y_i the target and X_i the data.

- For $\ell 1$ regularization we minimize the following cost function.

$$\min_{w,c} \parallel w \parallel_1 + C \sum_{i=1}^{n} \log(\exp(-y_i(X_i^T w + c)) + 1) \tag{2.13}$$

- $\ell 2$ regularization solves the following optimization problem.

$$\min_{w,c} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \log(\exp(-y_i(X_i^T w + c)) + 1) \tag{2.14}$$

- The Elastic-Net regularization (the combination of $\ell 1$ and $\ell 2$) optimizes

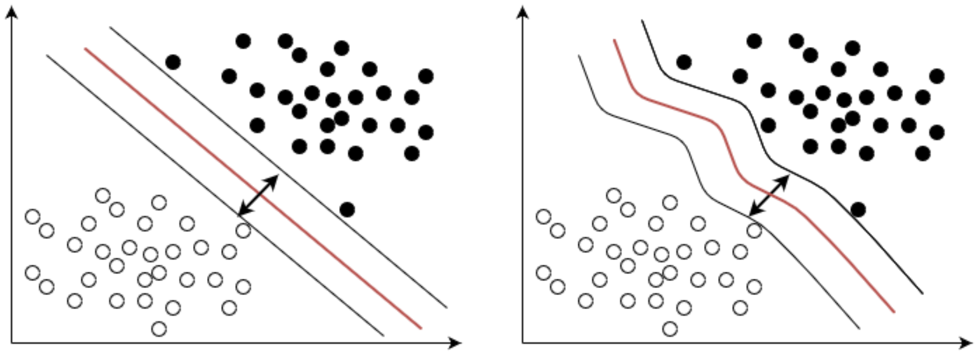
$$\min_{w,c} \frac{1-\rho}{2} w^T w + \rho \parallel w \parallel_1 + C \sum_{i=1}^{n} \log(\exp(-y_i(X_i^T w + c)) + 1) \tag{2.15}$$

where ρ determines the strength of ℓ_1 and ℓ_2 regularization.

2.5.2. Support Vector Machines

A Support Vector Machine (SVM) is a classification technique with the aim to maximize the margin between the class border and the samples closest to it (the support vectors). The borders are entirely determined by the support vectors and are either mapped in a linear or non-linear space [5]. An example of possible separation lines or hyper-planes is shown in Figure 2.3.

[2] `https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression`



(a) Linear separation. **(b)** Non-linear separation.

Figure 2.3: SVM separating the classes by maximizing the distance between the nearest points to the border (support vectors) and a) linear and b) non-linear class border.

For non-linear SVMs a small number of support vectors is desired, as it reduces the computational cost and the complexity of the resulting border, which decreases the risk of overfitting. To further increase the efficiency, the so called kernel trick is applied. The term describes the computation of the more efficient kernel function K instead of the mapping Φ, which is of high importance when working with real data. In practice it is most common to use either polynomial or Gaussian radial basis function (RBF) kernels. They are defined as follows

- **Polynomial kernels**

$$K = (\vec{x}, \vec{z}) = (\vec{x}^T \vec{z} + c)^P, c \geq 0 \tag{2.16}$$

  where c is the influence of higher-order versus lower-order terms in the polynomial. When $c = 0$, the kernel is called homogeneous. P is the degree of the polynomial.

- **RBF kernels**

$$K = \exp - \frac{\| \vec{x}_i - \vec{x} \|^2}{2\sigma^2} \tag{2.17}$$

  where σ determines the range of the "neighborhood" and thus the support vectors.

New data $\vec{x}$ is predicted according to the following decision rules. Where α_i are the Lagrange multipliers, y_i the targets and b the intercept of the border.

- Linear border

$$\sum_i \alpha_i y_i \vec{x}_i^T \vec{x} + b \geq 0 \rightarrow class\,1 \tag{2.18}$$

- Non-linear border with kernel trick

$$\sum_i \alpha_i y_i K(\vec{x}) + b \geq 0 \rightarrow class\,1 \tag{2.19}$$

- Special Case of 2.19: Non-linear border with RBF kernel trick

$$\sum_i \alpha_i y_i \exp - \frac{\| \vec{x}_i - \vec{x} \|^2}{2\sigma^2} + b \geq 0 \tag{2.20}$$

Generally, the data cannot be perfectly separated into the respective classes. This is where soft-margin SVM comes into place. A regularization parameter C is introduced, which determines whether data points are allowed to be placed in the margin or even in the other class. A large C highly penalizes any intrusion into the margin whereas a small C will prefer having a wide margin with the cost of misclassification. The optimization problem to be solved is as follows [5].

$$minimize \frac{1}{2} \| \vec{w} \|^2 + C \sum_i \zeta_i \tag{2.21}$$

with

$$\zeta_i = \max(0, y_i(1 - \vec{w}^{\,T} \vec{x}_i + b)) \tag{2.22}$$

and

$$\vec{w} = \sum_i \alpha_i y_i \vec{x}_i \tag{2.23}$$

where ζ_i is the penalty and w the normal to the border.

SVMs are effective for numerous applications. They work well in high dimensional spaces or in case the number of dimensions exceeds the number of samples. Moreover, they use a subset of the training data in the decision function (support vector), which makes it memory efficient. Additionally, it is possible to make use of different kernel functions as decision function [3].

[3] https://scikit-learn.org/stable/modules/svm.html#svm-classification

2.5.3. Random Forest

A Random Forest is a classification technique consisting of multiple independent, random decision trees. Decision trees start at the root node, that represents the whole data set. They are then built iteratively by asking several yes/no questions, making them therefore suitable for categorical classification tasks. The data is split in classes or attributes until a predetermined stop criterion is reached. Owing to its NP completeness, it is nearly impossible to find the optimal decision tree. Hence, practically algorithms are based on heuristic algorithms such as the greedy algorithm taking only the local optimum at each node into consideration [5]. For the creation or training of a decision tree several algorithms can be used. The algorithms differ only in their splitting strategy. The most common algorithms are the Hunt's algorithm, ID3 (Iterative Dichotomiser 3), C4.5 and CART(Classification and Regression Trees) [21]. The latter is utilised in the scikit-learn library which is used in this thesis. Figure 2.4 shows an exemplary of such a decision tree.

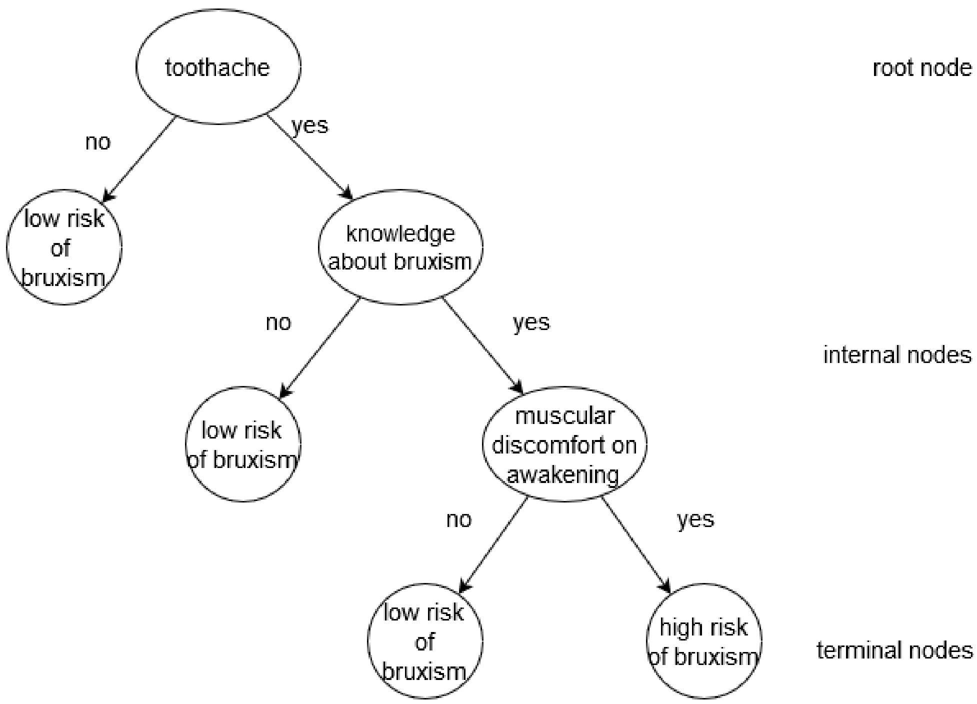


Figure 2.4: Example of a Decision Tree with three symptoms of bruxism.

There is also a risk of overfitting decision trees, because there is almost always another feature that can be used for splitting. To lower the risk, the complexity of the grown decision tree has to be reduced, e.g. applying one of the following techniques [5].

- **Pre-pruning:** Stop splitting if the information gain (Δ) is less than a threshold
- **Post-pruning:** Replace smaller sub-trees with leaves, e. g. if the number of samples is lower than a threshold
- **Minimum Description length (MDL) principle**: Select between trees of different complexity. The aim here is to minimize complexity.

Information gain can be calculated using the *Gini index* or Entropy criterion, which represent the impurity of a node. A node with only one class is considered a pure class. The existence of multiple classes make a node impure. Gini and entropy are defined as follows

$$Gini = 1 - \sum_{i=1}^{N} p^2(c_i) \tag{2.24}$$

$$Entropy = \sum_{i=1}^{N} -p^2(c_i) \log_2(p(c_i)) \tag{2.25}$$

where $p(c_i)$ is the probability of class c_i in a node.

Random Forest aims at building an ensemble of decision trees with high variance and low bias, which combined build a forest with low bias and low variance. This is done by:

1. Growing multiple trees from resampled training data.
2. Choosing only a subset of n out of N attributes ($n << N$) at each node for splitting
3. No pruning, i. e. allowing the risk of overfitting

New samples are predicted using a majority vote of the whole forest. The quality of the forest depends on whether the decision trees are independent (highly correlated trees will perform worse, because the prediction will be nearly the same) and the number of features (a small number cannot model the complexity of the data correctly) [5].

2.6. Evaluation of Classifiers

The quality of a classifier can be quantified using different metrics or scores. Typical metrics are accuracy, classification error, recall, precision and F1-score [23]. These can be derived from the so called confusion matrix. Table 2.1 shows the confusion matrix for a binary classification problem. The mathematical representation of the metrics are presented in Equation 2.26 to 2.30.

	predicted classes	
	positive	negative
positive	True Positive (TP)	False Negative (FN)
negative	False Positive (FP)	True Negative (TN)

Table 2.1.: Confusion Matrix for binary classification problem.

Although, accuracy and *error* are considered as standard metrics, they are not suitable for imbalanced data sets. They tend to classify according to the majority class, which can lead to wrong results and conclusions. As a result, the other scores (eq. 2.28 and 2.29) are preferred for applications where the performance of both classes (positive or negative samples) are equally important. The F1-score (eq. 2.30) enhances one of the classes, i. e. determines relevant elements. Another score used in machine learning is the ROC-AUC score. The metric computes the Area Under the Receiver Operating Characteristic Curve (ROC AUC). The ROC plots the fraction of TP out of the positives (TPR = true positive rate) against the fraction of FP out of the negatives (FPR = false positive rate) at different thresholds [4].

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{2.26}$$

$$Error = \frac{FP + FN}{TP + FN + TN + FP} \tag{2.27}$$

$$Recall = \frac{TP}{TP + FN} \tag{2.28}$$

$$Precision = \frac{TP}{TP + FP} \tag{2.29}$$

$$F1 - Score = 2 \cdot \frac{Precision \times Recall}{Precision + Recall} \tag{2.30}$$

Cross-Validation (CV) is a procedure applied to overcome overfitting by splitting the training set into training and validation set. k-*Fold* CV is the basic approach. The training set is split into k folds (subsets). The model is trained using $k-1$ of the folds as training data. The derived model is then validated with the remaining fold. The performance metric of the k-fold CV is the average of the computed validation scores per split.

[4] https://scikit-learn.org/stable/modules/model_evaluation.html#roc-metrics

2.7. Calibration of Classifiers

Calibration comes into place, when it is important to assure the generalizability of a classifier. Medical applications, as investigated in this thesis, are great examples of this. It is crucial for the classifier to accurately predict on a new subject or patient. There exist multiple methods for calibration, whereas *Platt Scaling* and Isotonic Regression are the most common. They work well for Logistic Regression as well as Support Vector Machines [4]. Literature on calibration methods for Random Forests is ambiguous. While some papers claim that they are relatively well calibrated, others acknowledge that they may be improved [6]. Relevant calibration techniques are described in the following.

2.7.1. Platt Scaling

Platt scaling, often referred to as *sigmoid scaling*, is a calibration method mostly applied on SVM. It consists of optimizing the parameters of a sigmoid function, such that the likelihood of the training set is maximized [1, 6]. The function is given by Equation 2.31.

$$\hat{p}(c|s) = \frac{1}{1 + \exp{(As + B)}} \tag{2.31}$$

Where $\hat{p}(c|s)$ corresponds to the probability that a sample is part of class c, given that it has achieved a score of s. A and B are the parameters of the function and are fitted using *gradient descent* to minimize

$$min\{-\sum_i y_i \log{(p_i)} + (1 - y_i) \log{(1 - p_i)}\} \tag{2.32}$$

where $p_i = \hat{p}(c|s)$.

2.7.2. Isotonic Regression

Isotonic Regression is a calibration technique that can be interpreted as "a general form of binning that does not require presetting of a specific number of bins or any limits of the size of each bin" [1]. The calibration function is supposed to be *isotonic*, i. e. non-decreasing. Given f_i, the predictions of the applied classifier, and the true labels y_i, the isotonic regression is

$$y_i = m(f_i) + \epsilon_i, \tag{2.33}$$

where m is an isotonic function and ϵ is an observation error. Given a training set (f_i, y_i), where f_i is the prediction made by the model and y_i the true targets. The objective of the Isotonic Regression is to find the non-decreasing function $\hat{m}$ as follows.

$$\hat{m} = \arg\min_{z} \sum (y_i - z(f_i))^2 \tag{2.34}$$

This increases not only the flexibility, but also the risk to overfit. Hence, Isotonic regression yields better performance than Platt scaling only if there is enough data to avoid overfitting [6].

2.7.3. Calibration of Random Forests

In [1], a method based on the finding that the squared error of probabilities predicted by forests of classification trees are lower compared to those predicted by forests of probability estimation trees (PET), is presented. The technique consists of correcting the averaged probability distribution. The estimated probability for the most probable class is increased by parametrizing the increase as follows:

$$\hat{p}_i = \begin{cases} p_i + r(1 - p_i) & \text{if } p_i = \max\{p_1, \ldots, p_k\}, \\ p_i(1 - r) & \text{otherwise} \end{cases} \tag{2.35}$$

where $p_1, \ldots p_k$ are the uncorrected probability estimates for the k classes and $r,\ 0 \leq r \leq 1$, is the correction parameter that pushes the estimated probability of the most probable class to a specific value by reducing the other probabilities. For example, pushing it to 1 with $r = 1$ will push the other classes to 0.

Dankoswki and Ziegler [4] propose a method where the random forest is first transformed into logistic regression models. This is done by estimating the relative frequencies of all subjects having the event in the terminal node. Logistic regression can be used to estimate the conditional probabilities. Let $P(y = 1|x_1 \leq c_1, x_2 \leq c_2)$ be the conditional probability at a terminal node t, then the dummy variable is defined as follows

$$d_1 = \begin{cases} 1 & \text{if } x_1 \leq c_1, x_2 \leq c_2 \\ 0 & \text{else} \end{cases} \tag{2.36}$$

Dummy variables are defined for each terminal node. The one with the largest amount of samples in the terminal node is chosen as reference category. A logistic regression model is fitted for every former tree. In case that the tree consists of three terminal nodes, the model to be fitted is

$$P(y = 1|d_1, d_2) = \alpha + \beta_1 d_1 + \beta_2 d_2 \tag{2.37}$$

if d_3 is the reference category. This is done for all trees in the forest. The random forest is then updated using re-calibration of each logistic regression model.

2.7.4. Evaluation of the Calibration

In general, the performance of probability estimation models or calibration is measured using the *Brier score* (BS). The score is defined as the mean-squared difference between the patient status and the predicted probability [4].

$$B = \frac{1}{N} \sum_{i=1}^{N} (\hat{p}_i - Y_i)^2, \tag{2.38}$$

Where $\hat{p}_i$ is the estimated probability of the observation i and Y_i is the observed value. The calibration is considered to be better, the smaller the Brier score.

3. Materials and Methodology

The previous chapters focused on introducing the main concepts necessary for the design of a biofeedback system. An overview on the functionality of electromyography, the importance of feature extraction and the selection of classification techniques were presented. In this chapter the preliminaries to develop a concept for a EMG biofeedback prototype are presented. The first step consists of the conduction of several experiments as described in Section 3.3, to explore EMG data and to examine the previously determined research questions. The results will be presented in Chapter 4 and summarized in the section "Conception of an EMG-Biofeedback System".

3.1. Materials

The main components utilized for the conduction of the experiments are described in the following. An EMG sensor with surface electrodes connected to a microcontroller has been used for data acquisition.

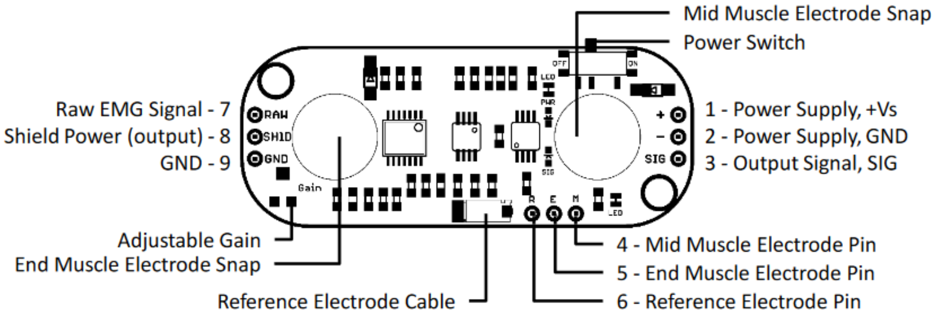
3.1.1. EMG Sensor

To obtain the EMG data I used the *MyoWare Muscle Sensor (AT-04-001)*, an electromyography sensor for microcontroller applications. As a result of its wearable design data acquisition was facilitated. The sensor features several input and output pins, two electrode snaps and a reference electrode cable (see Figure 3.1a). It provides two output types, the raw EMG signal (Figure 3.1a (7) RAW) and the EMG Envelope (Figure 3.1a (3) SIG). In this work, the latter was used, which is in essence the amplified, rectified and integrated signal and thus suitable for the analog-to-digital converter (ADC) of a microcontroller [1]. Figure 3.1b depicts the difference between the output types.

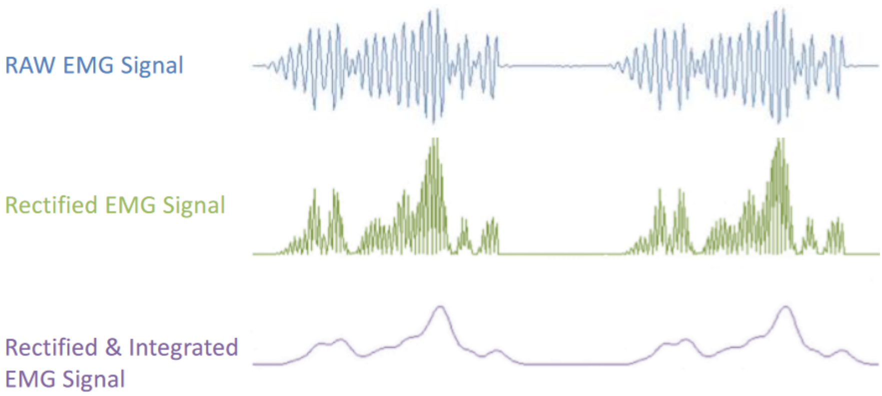
3.1.2. Electrodes

For every recording three electrodes were required. One for the mid muscle, one for the end muscle and the last for the reference bone or muscle (compare with Figure 3.1a). The sensor kit used in this thesis came along with Kendall™ ECG Electrodes Product Data SheetArbo™ H124SG electrodes. The electrodes have a sensor area of $80\,mm^2$

[1] https://cdn.sparkfun.com/assets/learn_tutorials/4/9/1/MyoWareDatasheet.pdf



(a) Layout of the MyoWare Muscle Sensor AT-04-001



(b) Difference between RAW EMG Signal and EMG Envelope

Figure 3.1: Characteristics of the MyoWare Muscle Sensor Source: [1]

which maximizes the electrode distance that should range between 10 and $40\,mm$. As stated in Section 2.1.2 the use of a conductive gel and silver-chloride electrodes were recommended for noise reduction. These requirements are met by the electrodes [2]. This promises good recordings with low noise, which has been confirmed by initial experiments and tests.

3.1.3. Microcontroller

As mentioned before, the sensor is designed for microcontroller applications. For this work an Arduino Leonardo board, that is based on the 8-bit AVR microchip ATmega32u4, has been chosen. The board features various input/output pins and a micro USB connection. The latter is used to connect the device with the computer for data collection and setup. The former are used to connect the sensor, the actuator and the battery. The test setup will be presented in Section 3.2.3.

3.1.4. Vibration Motor

A $5\,V$ vibration motor suitable for Arduino has been used as an actuator for the biofeedback device. The module has three pins (IN, VDD, GND). The input pin (IN) can be connected to one of the PWM pins of the Arduino. Figure 3.2 shows a depiction of the module.

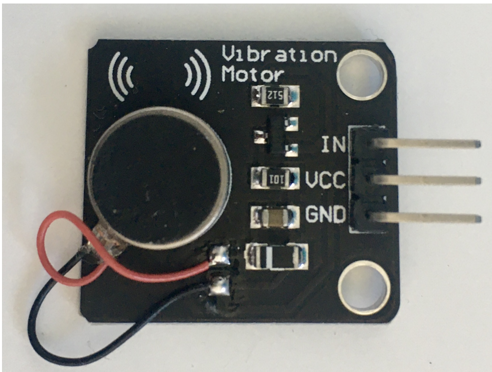


Figure 3.2: Figure of the vibration motor module used in section 4.4.

3.2. Data Acquisition

Data is acquired from three subjects. This section describes how and where the data is collected.

[2] https://www.mouser.com/datasheet/2/813/H124SG-1022817.pdf

3.2.1. Jaw Muscles Activity while Sleeping (Scenarios)

It is known that human perform multiple jaw muscle activities throughout the day. However, their activities are not only restricted to the time of day. In the following section the most common jaw muscles activities that occur while sleeping (oral parasomnias) are listed. The list is used to derive a collection of movements to be recorded during data acquisition.

- The subject sleeps calmly without activation of the jaw muscles. (resting)
- The subject yawns while sleeping. (yawning)
- The subject clenches his/her teeth while sleeping. (teeth clenching)
- The subject grinds his/her teeth while sleeping. (teeth grinding)
- The subject chews while sleeping. (chewing)
- The subject swallows while sleeping. (swallowing)
- The subject talks while sleeping. (talking)
- The subject makes other movements while sleeping.

For this study I decided to opt for the following movements: resting, yawning, teeth clenching, teeth grinding and chewing.

3.2.2. Muscle Choice

The performance of jaw activities is related to different muscles. However, the masseter and the temporalis muscle are most commonly used for surface EMG recording as they are easily accessible [9]. Figure 3.3 depicts their location [3].

Table 3.1 shows that the data acquired at the temporalis muscle has a higher distribution than the masseter muscle. This can be seen on the standard deviation of the recorded signals for grinding and clenching, calculated using the *DataFrame* structure from the Python library *Pandas*.

$$\sigma_{grinding,temporalis} = 98.8905 > \sigma_{grinding,masseter} = 14.3937$$

or

$$\sigma_{clenching,temporalis} = 26.8950 > \sigma_{clenching,masseter} = 15.3723$$

This relation is also true for other movements than yawning. Since this thesis focuses on the detection of teeth grinding and clenching, 'yawning' is negligible. Moreover,

[3] `http://what-when-how.com/dental-anatomy-physiology-and-occlusion/the-temporomandibular-joints-teeth-and-muscles-and-their-functions-dental-anatomy-physiology-and-occlusion-part-3/`

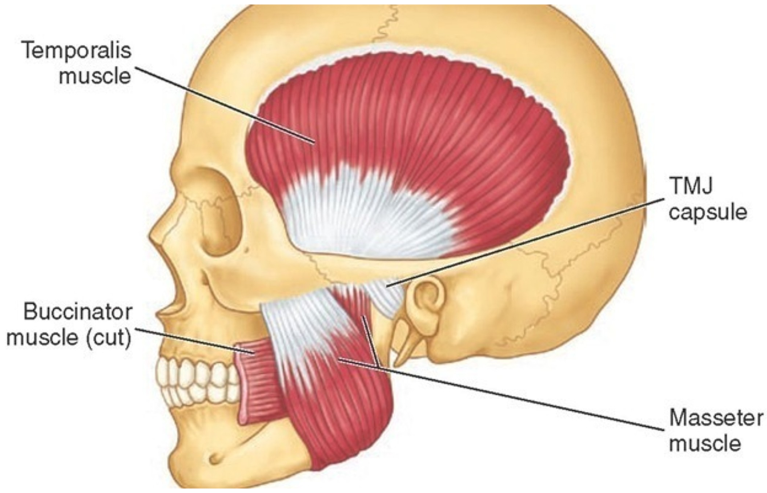


Figure 3.3: Location of Masseter and Temporalis Muscle. Source: [3]

the max values, describing the intensity of the muscle activation, show that the EMG device is more sensitive at the temporalis muscle than at the masseter muscle.

$$max_{grinding,temporalis} = 453.0 > max_{grinding,masseter} = 93.0$$

and

$$max_{clenching,temporalis} = 195.0 > max_{clenching,masseter} = 111.0$$

For clenching no significant difference can be obtained, however $25\ \%$ of the temporalis data is above 55 which is $\Delta = 9$ unities more compared to the data measured at the masseter muscle. Additionally, placing the EMG sensor at the temple does not require any further measures e. g. shaving in case of a placement at masseter muscle. Thus, all the following measurements will only be conducted on the temporalis muscle.

3.2.3. Test Setup

For the acquisition of the EMG data a computer, an Arduino Leonardo and the MyoWare muscle sensor were used. The test setup is depicted in Figure 3.4. Some parts of the figure were taken from [4]

The sensor is connected to the person by placing the **M** electrode to the mid, the **E** electrode to the end of the respective muscle and the **R** electrode to a bony area as far away as possible from the muscle. + and – are connected to the $5V$ and GND pins of the Arduino respectively. The signal pin (SIG) of the sensor is connected to one of the analog pins of the Leonardo, here: $A0$. The board itself is connected to a computer via micro USB. The computer can be replaced by any device able to run Python scripts. The script used for data acquisition is described in Section 3.2.4.

[4] `https://store.arduino.cc/arduino-leonardo-with-headers`

	resting	**yawning**	**chewing**	**clenching**	**grinding**
count	55 334	54 718	55 045	54 771	55 334
mean	10.538	19.664	21.448	30.981	38.609
std	0.508	41.924	16.433	14.394	15.372
min	9	9	9	9	9
25%	10	11	11	21	28
50%	11	11	14	27	36
75%	11	12	25	38	46
max	12	548	135	93	111

(a) Masseter Muscle.

	resting	**yawning**	**chewing**	**clenching**	**grinding**
count	55 043	54 534	55 041	54 988	55041
mean	22.501	41.535	34.007	98.900	40.538
std	29.150	40.652	26.492	98.900	26.895
min	8	8	9	9	8
25%	9	13	16	19	20
50%	11	28	26	59	34
75%	17	57	44	159	55
max	193	357	310	453	195

(b) Temporalis Muscle.

Table 3.1.: Statistics describing the distribution of the data acquired from both, masseter and temporalis muscle.

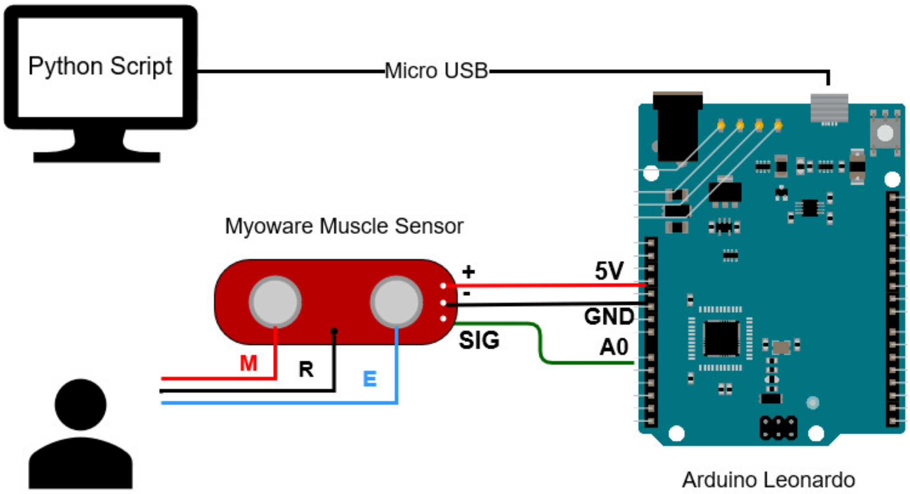


Figure 3.4: Block diagram of the test setup. Parts from [4]

3.2.4. Data Collection Procedure

The EMG data has been collected performing the following steps.

1. **Connect Arduino to computer.**

   The board is connected according to the test setup presented in Section 3.2.3. The Arduino Sketch for communication is already batched.

2. **Find the right position to place the Electrodes.**

   The subject is asked to tense the jaw muscle in order to locate the temporalis muscle.

3. **Start the script.**

   a) Select the right port and choose a file name (optionally).

   b) Perform the following tasks sequentially.

      i. Grind your teeth for $4\,s$ straight. Relax for another $4\,s$. Repeat five times.

      ii. Clench your teeth for $4\,s$ straight. Relax for another $4\,s$. Repeat five times.

      iii. Chew for $4\,s$ straight. Relax for another $4\,s$. Repeat five times.

      iv. Yawn for $4\,s$ straight. Relax for another $4\,s$. Repeat five times.

v. Rest for $60\,s$.

An Excel File containing the time series for each movement was received. The data is of numerical type. The values range from theoretically 0 to 1023 which is the output range of the Arduino. The exact value range depends on the subject and the amplitude of his or her masticatory activity.

3.3. Experimental Design

The experiments are divided into three categories.

1. **Experiments of Type 1**

   The first experiments consist of pre-analysis and the familiarization with the EMG sensor device and its output. The objective is to obtain a workflow for EMG data acquisition.

2. **Experiments of Type 2**

   The second type of experiments focus on the analysis of EMG data from subject 1. The acquired data is used as reference data. It is analyzed, processed and then fed to the classifiers.

3. **Experiments of Type 3**

   Type 3 contains experiments to check whether classifier calibration can be used to generalize the classifier obtained from training with **only** subject 1. It is validated with data acquired from subject 2 and 3.

3.3.1. Experiment 1: Data Pre-Analysis

The experiments are aimed at getting an impression of whether the individual movements can be differentiated. This is done by

a) visually analyzing the different patterns for each recorded movement. The intention is to identify differences and similarities in shape, period and value range (intensity) of the time series representing the EMG signals.

b) performing a cluster analysis. In order to do so, the dimensionality of the sampled data needs to be reduced first. Therefore, the different time series are brought to equal length by cutting off samples exceeding the length of the recording with the lowest amount of samples. The Principal Component Analysis (PCA) explained in the previous chapter has been used for dimensionality reduction. The goal is to verify whether the assumption, that the problem can be formulated as a binary classification problem, is correct. The results will be presented in Section 4.1.

3.3.2. Experiment 2: Signal Processing and Data Analysis

Experiment 2 can be divided into several sub-experiments.

a) Data Preprocessing

As mentioned in section 3.2.4, data is stored in an Excel file. The file is composed of six spreadsheets containing the time series (EMG signals) for each movement. The Excel file is loaded into a DataFrame (data structure from the Python library Pandas) of size: $6 \times N$, where N is the length of the shortest time series. Samples exceeding this length are cut off.

Noise Cancelling

Surface EMG signals range from 0 to $500\,Hz$, whereby frequencies between 0 and $20\,Hz$ are unstable. In order to cancel these noises two Butterworth filter, i. e. a high-pass and a low-pass filter, were applied. This has been done through the functions *butter* and lfilter from the python library *Scipy Signal*. The *butter-method* accepts various input parameters, which were set as follows.

- N is the order of the filter. The common value for EMG applications was chosen, which is 6.
- Wn is the critical frequency that needs to be eliminated. It is defined as

  $$Wn = \frac{2 \cdot f}{f_s},$$

  where f is the frequency to be eliminated (here: $20\,Hz$ for the low-pass and $500\,Hz$ for the high-pass) and f_s is the sampling frequency.
- The filter type can be set to *'high'* or *'low'* using the parameter $btype$.
- The sampling frequency is described by fs. The typical sampling frequency is $1\,kHz$ and is also applied in this thesis.

The method returns the numerator and the denominator polynomials of the IIR filter that are used by the *lfilter-method* to compute the filtered signal. The filtered signal is stored as a dictionary with the movement as keys and the filtered signal in list format as values.

Framing and Labelling

The next step is to rearrange the signal by dividing it into frames of equal size. The frame size is determined by the frame duration and the sampling frequency.

$$frame\_size = frame\_duration \cdot sampling\_frequency$$

In this thesis a frame duration of $3\,s \approx 3144$ samples was chosen, which corresponds to the typical duration of nocturnal teeth grinding or clenching. Due to the high number of noise in the recorded EMG curves, the frames were selected and labelled manually. This was done by joining the individual recordings and detecting the significant peaks of the respective movements. Signals that lasted longer than $3\,s$ were covered using overlapping, i. e. a signal of $5\,s$ ranging from e. g. $150\,k$ to $155\,k$ was represented by two frames from $150\,k - 153.1\,k$ and $152\,k - 155\,k$. The values listed before are the sample numbers. Frames corresponding to the movements *grinding* and *clenching* were labelled as 1. *Chewing, yawning* and *resting* were labelled as 0.

b) Feature Selection and Extraction

The following sub-experiments are closely connected to the sub-experiments c). The aim is to verify how classification performance changes when classifying either with or without feature selection. In the time domain, features presented in Section 2.2.1 are examined. They were implemented in Python. For the time-frequency domain features the method *dwt* from the *pywt* library was used.

The function returns the approximation and detail coefficients, i. e. the features. It takes the following input arguments.

- The $data$ to process,
- The $wavelet$-function to apply. Here the *Daubechie* function was chosen, as it has proven its application for surface EMG.
- And finally, the signal extrapolation $mode$, which was set to the 'constant' mode, where borders are replicated.

c) Training and Evaluation of Classifiers

In this thesis the focus is on three classifiers of different types. A linear classification model was applied, the **Logistic Regression**, a **Support Vector Machine** and a **Random Forest Classifier**. Before training the data set was split into a training and testing set using the $sklearn.model\_selection.train\_test\_split$ function. A splitting ratio of $1:3$ was chosen.

For training the processed training data was used. The processing steps described in the above paragraphs were followed. To obtain the best parameters the *GridSearchCV* function was used, which optimizes the parameter of an estimator applying a cross-validated grid-search over a parameter-grid [5]. The function requires the estimator and the parameter-grid as input arguments. Additionally, it takes optional input arguments. For this thesis only the number of folds for cross-validation (cv) and the $scoring$-function are of interest. Considering, that the classifier should be reliable when detecting "bruxism", the 'recall'-score was chosen as objective function.

The parameter-grids used for each classifier are shown below.

- **Logistic Regression**

  The norm used for penalization is defined by using the term $penalty$. The parameter C specifies the regularization. A smaller value specifies a stronger regularization. The weights associated with each class are defined by $class\_weight$. If $None$ is given, all classes have weight one. If $balanced$ is chosen, it "uses the values of y to automatically adjust weights inversely proportional to class frequencies in the input data as"

  $$\frac{N}{M \cdot b_y} \tag{3.1}$$

  where N is the number of samples, M the number of classes and b_y the number of occurrences of each value in an array of Integers [6].

  Parameter Grid
  The combination of the following parameters are tried
  - ℓ_1 and ℓ_2 penalty
  - Regularization value C of $0.1,\ 1$ and 10
  - Class weights of 1 for each class, 2 for Class 1 and 1 for Class 0 as well as balanced class weights

- **Support Vector Machine**

  The parameters for the linear SVM are the same as for the Logistic Regression. For the non-linear SVM a kernel-function using the keyword $kernel$ can additionally be determined. The γ-parameter is the kernel-coefficient used for *'poly',*

[5] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html?highlight=gridsearch%20cv%23sklearn.model_selection.GridSearchCV

[6] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html?highlight=logistic%20regression#sklearn.linear_model.LogisticRegression

'rbf and 'sigmoid'[7]. It can either be a given float or the keywords *'scale'* and *'auto'*. If γ is set to 'scale', which is the default value,

$$\gamma = \frac{1}{N * \sigma^2(X)} \tag{3.2}$$

is used as value of gamma. Where N is the number of samples and σ^2 the variance of the signal X. 'auto' uses

$$\gamma = \frac{1}{N} \tag{3.3}$$

Parameter Grid
For the linear SVM the following parameters were tried.

- ℓ_1 and ℓ_2 penalty
- Regularization value C of $0.1,\ 10$ and 100
- Class weights of 1 for each class and balanced class weights

For the non-linear SVM the parameter grid looked as follows.

- Polynomial, RBF and sigmoid kernels
- Regularization value C of $0.1,\ 10$ and 100
- Gamma values γ of 0.01 and 10
- Class weights of 1 for each class and balanced class weights

• **Random Forest Classifier**

The Random Forest Classifier parameters tested are shown below.

- The number of trees in the forest $n\_estimator = [10,\ 100\,1000]$
- The splitting criterion, the *Gini index* or entropy
- The minimum number of samples required to split an internal node $[0.5,\ 2,\ 10]$
- The minimum number of samples required at a leaf node $[2,\ 5,\ 12]$
- Class weights of 1 for each class and balanced class weights

It is common to apply dimensionality reduction before classifying. In order to investigate the necessity in the application, the best estimator was fitted with and without PCA.

[7] `https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html?highlight=svc#sklearn.svm.SVC`

3.3.3. Experiment 3: Calibration

As mentioned previously, surface EMG data is non-stationary. The amplitude of the EMG signal varies for different persons. They also change for one person during the day. It is thus necessary to calibrate the prediction model before performing classification, in order to improve the detection performance. This section contains a comparison of the calibration techniques presented in Section 2.7. The standard methods Plott Scaling and Isotonic Regression will be evaluated. Calibration will be tested with data acquired from subject 2 and 3.

Platt Scaling and Isotonic Regression

Calibration was done using the CalibratedClassifierCV function from the Scikit Learn *calibration* library. The function applies Platt Scaling and Isotonic regression on the model with the best score presented in Section 4.2. Calibration is validated using 5-fold cross validation.

Evaluation with Brier Score

The Brier Score has been obtained using the brier_score_loss function from the Scikit Learn *metrics* library. The probabilities are calculated using the classifier method *predict_proba*. The score is computed from the true labels and the probabilities of the positive class for each label. It is interpreted as follows. The lower the Brier score, the better the calibration performance.

Data Preparation

Unlike for the training of the model, as described in Section 3.3.2, calibration data is not framed and labelled manually. It is done by automatically dividing the recorded signals into frames of equal size. The frame size is determined by the frame duration and the sampling frequency. In this thesis a frame duration of $3\,s$, the typical duration of nocturnal teeth grinding or clenching, is chosen. In order to increase the number of samples, the windows are overlapped. Each half of the window appears in two frames.

In this work a binary classification problem has been defined. The positive class contains the movements 'grinding' and 'clenching', the negative one contains the other movements. 100 frames, or precisely the features extracted from them, were utilized to fit the calibrated classifier. The remaining 60 and 70 samples for Subject 1 and Subject 2 respectively, were used to evaluate the calibration using the Brier Score.

4. Results

In Chapter 3 several experiments were conducted to investigate the defined research questions. The experiments have been divided into three categories, each with different objectives. First, pre-analysis has been done to evaluate whether the different movements were generally distinguishable. Second, the performance of the machine learning algorithms SVM, Random Forest and Logistic Regression were examined. Lastly, classifier calibration using Platt Scaling and Isotonic Regression was tested on two further subjects. The obtained results are presented in this chapter.

4.1. Experiment 1: Data Pre-Analysis

Pre-Analysis consisted of a visual analysis of early recordings on the reference subject and the execution of a cluster analysis using DBSCAN. The following findings were made.

4.1.1. Visual Analysis of the Recordings

Figure 4.1 shows an excerpt from an EMG signal used to identify similarities and differences between the jaw movements. By roughly examining the excerpt, we can detect one movement that particularly strikes out, which is the red curve. It depicts the unprocessed signal for 'clenching'. The graph shows, that clenching has the highest intensity among all jaw movements. Contrary to the others, it is an aperiodic signal. The pulse width is around 1000 samples, which corresponds approximately to $1\ s$.

The blue curve shows the signal recorded while 'resting'. This is considered the reference signal for noise. There is a low amplitude showing that ECG artifacts are not a concern for the task at hand. Moreover, the amplitude is almost always lower than the other amplitudes, showing that the recorded signals are valid, i. e. that data is being recorded.

At first glance, the other signals have similar patterns. They are periodic and have about the same amplitudes. However, they have different periods. While 'chewing' and 'grinding' appear to be square signals, 'yawning' appears to be of sine-shape. The similarity between chewing and grinding can be explained by the similar execution of both jaw movements.

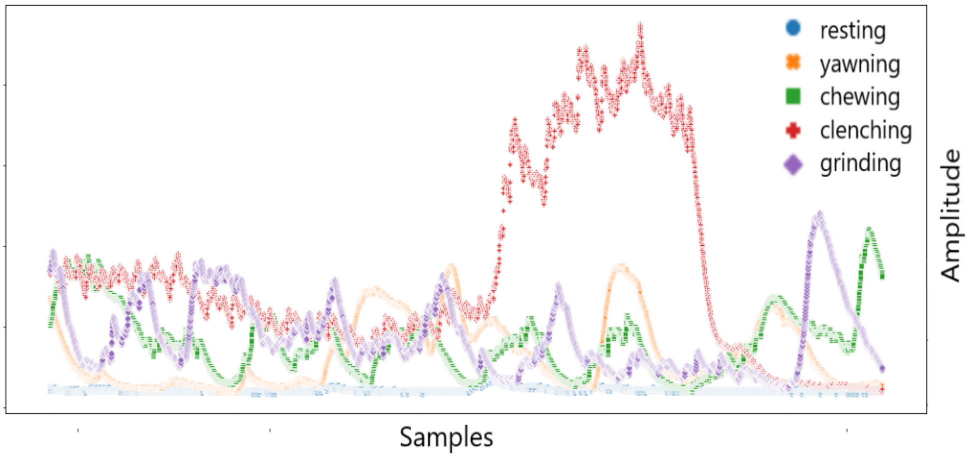


Figure 4.1: Excerpt from the EMG Signal

4.1.2. Cluster Analysis

The classification of a disease is usually a binary classification problem – whether it is positive (e. g. bruxism) or not (e. g. no bruxism). In order to check whether this classification type was also applicable for the data set as presented in this thesis, clustering on the pre-processed data with the DBSCAN algorithm has been performed. The resulting classes were compared to the real labels. Figure 4.2 shows the results of the DBSCAN with a maximum distance between two samples of $eps = 0.4$ and the minimum number of samples required for a point to be considered a core point of $min\_samples = 8$ (in the left) and the true labels on the right side. It can be seen that the samples of class 0 (*grinding and clenching*) are separable from those of class 1 containing *chewing, yawning and resting*. However, as expected, there is a small error when compared to the actual labels.

4.2. Experiment 2: Signal Processing and Data Analysis

This section contains a presentation of the results obtained from the data analysis made on the reference subject. In order to classify the quantitative results qualitatively, the data structure is first introduced. Moreover, the significant patterns of the EMG signals for each movement are analysed. Finally, the performances of the classifiers are presented and compared to each other.

4.2.1. Data Set

Table 4.1 shows the structure of the data used for training and testing of the different classifiers. Data has been acquired during 7 recording sessions, resulting in a data

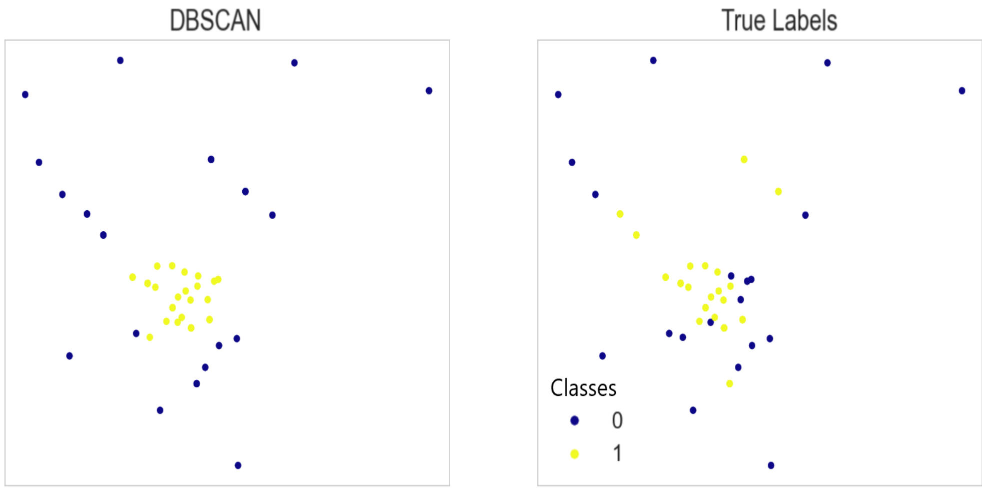


Figure 4.2: True labels compared with labels obtained from DBSCAN on PCA-reduced data

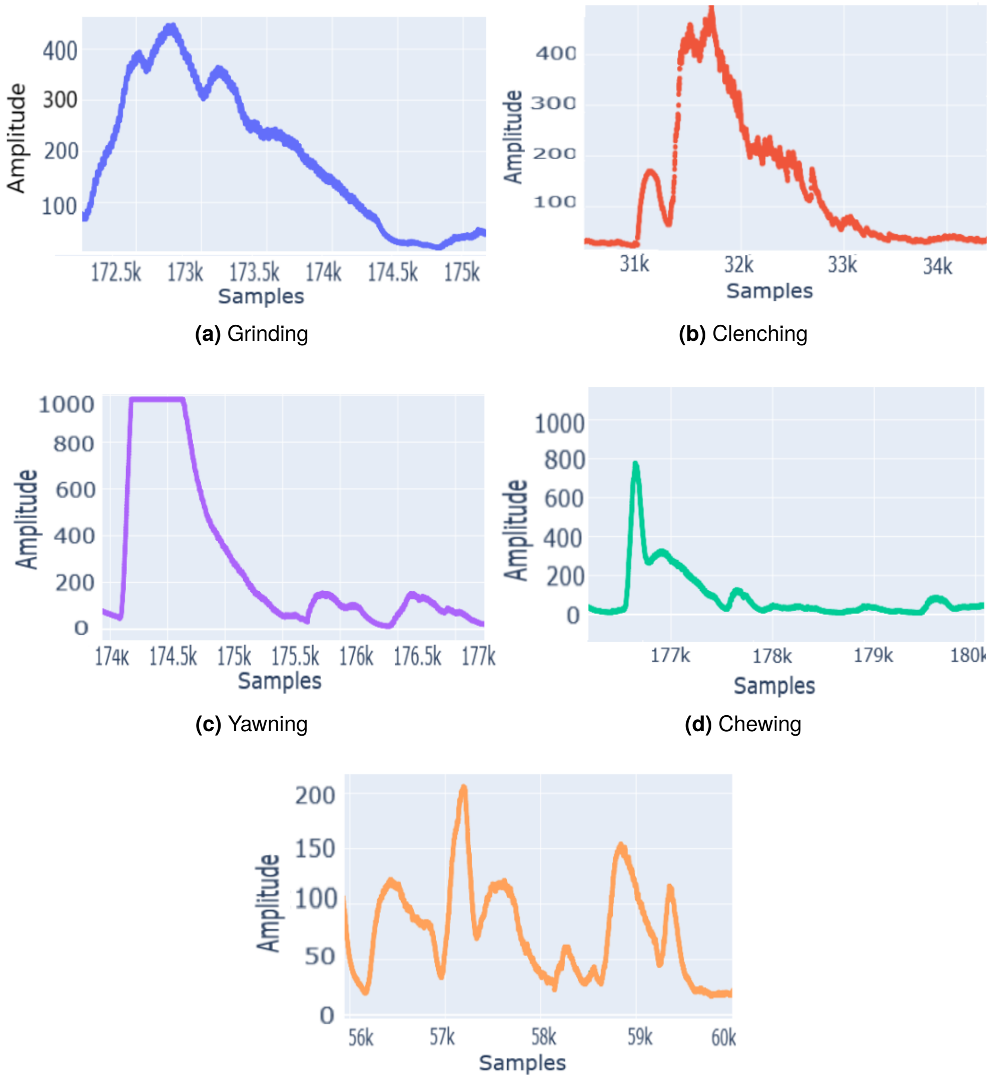
set of shape $(384394, 5)$. Where the former is the number of samples and the latter corresponds to the number of movements. After framing and labelling according to the method described in the section on a) Data Preprocessing, 88 frames – 44 for each label – are obtained. Each frame is of length 3144, representing a signal of approximately $3\,s$. Training and testing is done with a $33\,\%$ split. Hence, 58 frames are used for training and 30 for testing, respectively.

Number of Recordings	7
Train, Test Split Ratio	3:1
Shape of Training Data	(58, 3144)
Shape of Testing Data	(30, 3144)
Frame Duration [s]	3
Frame Duration [samples]	3144
Number of Frames	88
Labels	0: resting, chewing, yawning 1: grinding, clenching

Table 4.1.: Data Set Structure

4.2.2. EMG Signals

EMG signals of five different jaw movements have been recorded in the course of this work. The following figures show the unprocessed EMG signals, that are rectified, amplified and integrated by the sensor (see Section 3.1.1). The frames of the respective movements are depicted in Figure 4.1, illustrating the significant trends of each signal type.



(a) Grinding **(b)** Clenching

(c) Yawning **(d)** Chewing

(e) Resting

Figure 4.3: Excerpt from EMG Signal Recordings of each Jaw Movement.

Different patterns can be perceived from visual analysis. The depicted signal for **grinding** lasts for approximately $3000\,samples$. It has three consecutive peaks at $172.5k$, $172.8k$ and $173.2k$ with amplitudes of about 400, 450 and 360 respectively. The maximal achievable amplitude is 1023, which is related to the output range of the Arduino $(0 \ldots 1023)$. The area under the curve (AUC) is greater than for clenching and resting, but smaller than for yawning and chewing. Moreover, its main peak has the largest width ($1500\,samples$) amongst all.

Clenching has a less distinctive peak. It starts with a parabolic section with an amplitude of 180 and increases to almost 500 after approximately $1000\,samples$ or $1\,s$. It then decreases roughly linearly. The width of the peak is about $800\,samples$.

The highest amplitude, with 1000, is achieved through **yawning**. This value remains constant for circa $1\,s$, making the AUC for yawning the highest. The peak's width is similar to clenching. However, the actual signal or movement does not last for $3\,s$ but rather for $1.6\,s$.

Chewing has the smallest width of approximately $300\,samples$. It reaches an amplitude of 800.

The **resting** signal depicted in this frame is very noisy. The amplitude ranges from 20 to 210. The AUC of this signal is nevertheless the smallest.

Signal Amplitudes

The figures described above where only representative frames. However, there are some deviations in the pattern of each signal type. The signals primarily differ in their amplitude values. Table 4.2 summarizes the maximum amplitude $\hat{V}_{max}$, the average maximum amplitude per frame $\overline{\hat{V}}$ and the mean amplitude overall $\overline{V}$, for each movement. The amplitudes are described in $Volts$. Conversion is done using the following equation, where $5\,V$ is the operation voltage and $sEMG$ the value obtained by the sensor. The division with 1024 is relates to the output range of the Arduino.

$$V = sEMG \cdot \frac{5V}{1024} \tag{4.1}$$

Jaw Movement	$\hat{V}_{max}\ [V]$	$\overline{\hat{V}}\ [V]$	$\overline{V}\ [V]$
Grinding	4.228516	2.067755	0.874313
Clenching	3.076172	2.256284	1.025473
Chewing	3.789062	1.769456	0.442737
Yawning	4.799805	2.508042	0.795651
Resting	1.005859	0.483398	0.170472

Table 4.2.: Maximum and average amplitude of EMG signal for each jaw movement.

The table confirms that the highest peak with $4.8\,V$ is achieved yawning. It is closely followed by grinding ($4.23\,V$). Chewing presents the third highest amplitude with $3.79\,V$. Clenching is around $3\,V$. The resting signal achieves a maximum of $1\,V$. The average peak values however change the ranking to yawning, clenching, grinding, chewing and resting, in descending order. The mean values per signal are even lower for the movements labelled as '0' (no bruxism) as they contain more noise. This is due to the fact, that the frame size was chosen according to the duration of teeth grinding and clenching, which does not correspond to the signal length of yawning and chewing.

4.2.3. Logistic Regression

Parameters

As previously mentioned, different feature sets for training were tried. The model was trained with the raw data obtained from the sensor – in the following referred to as **None** –, the different time-domain features presented in Section 2.2.1, their combinations and the DWT coefficients retrieved according to Section 3.3.2 b). The highest score during training was achieved using a classifier with the parameters listed below. The results obtained while testing the trained models are summarized in Table 4.3.

- Regularization parameter $C = 0.1$
- Balanced class weights
- Maximum 10000 iteration steps allowing the classifier to converge
- ℓ_2 penalty

Performance

Logistic Regression performed on the raw signal provides overall a low performance. An accuracy of $53.33\,\%$ and a recall of $58.82\,\%$ are achieved. The application of time-domain features slightly increases the accuracy value to $56.67\,\%$. The recall increases to the maximum ($100\,\%$). The time-frequency domain feature DWT delivers the worst performance with values between $41.18\,\%$ $(recall)$ and $50\,\%$ $(precision)$. The accuracy is at $43\,\%$. However, performance can be increased applying PCA for dimensionality reduction. For $n\_components = 5, 10, 15$ an accuracy of $80\,\%$ and a recall value of $100\,\%$ was reached.

4.2.4. Support Vector Machine

Parameters

Highest training score with Support Vector Classification was achieved with the following parameters.

- The strength of regularization was set to $C = 0.1$.
- Classes are considered to be equally important, class weights are balanced.

	PCA	Training Score	Test Score				
Feature(s)	**N**	**Accuracy [%]**	**Accuracy [%]**	**Precision [%]**	**Recall [%]**	**F1 [%]**	**ROC-AUC [%]**
None	–	86.21	53.33	58.82	58.82	58.82	52.49
None	5	77.59	**80**	**73.91**	**100**	**85**	**76.92**
None	10	77.59	80	73.91	100	85	76.92
None	15	77.59	80	73.91	100	85	76.92
IEMG	–	46.55	56.67	56.67	100	72.34	50
SSI	–	46.55	56.67	56.67	100	72.34	50
Var	–	46.55	56.67	56.67	100	72.34	50
RMS	–	46.55	56.67	56.67	100	72.34	50
WL	–	46.55	56.67	56.67	100	72.34	50
LOG	–	46.55	56.67	56.67	100	72.34	50
SSC	–	46.55	56.67	56.67	100	72.34	50
RMS, SSC, WL, Var	–	46.55	56.67	56.67	100	72.34	50
IEMG, SSI, Var, RMS, WL, LOG	–	46.55	56.67	56.67	100	72.34	50
IEMG, SSI, Var, RMS, WL, WAMP, LOG, SSC	–	46.55	56.67	56.67	100	72.34	50
DWT	–	**89.66**	43	50	41.18	45.16	43.37

Table 4.3.: Test scores of the Logistic Regression model obtained for different feature sets.

- A polynomial kernel function with $degree = 4$ was chosen.
- The best suitable kernel coefficient was $\gamma = 0.01$.

Table 4.4 summarizes the testing results received with this classifier.

Performance

Classification with SVM gives the following results. The model trained with the unprocessed data provides an accuracy of $80\,\%$ and a recall and precision of $82.35\,\%$, which means that the probability of correctly classifying a signal is at $82.35\,\%$. The results are the same for DWT. Dimensionality reduction with PCA reduces the test scores to $accuracy = precision = 56.67\,\%$. Recall however increases to $100\,\%$, indicating that all positive samples are labelled as such ($FN = 0$). The same scores are gotten from IEMG, SSI, the combination of IEMG, SSI, Var, RMS, WL, LOG and the combination of all time-domain features. The best test score is achieved training a SVM model with the Waveform Length representing the complexity of the waveform in each segment.

	PCA	Training Score	Test Score				
Feature(s)	**N**	**Accuracy [%]**	**Accuracy [%]**	**Precision [%]**	**Recall [%]**	**F1 [%]**	**ROC-AUC [%]**
None	–	**100**	80	82.35	82.35	82.35	79.63
None	5	46.55	56.67	56.67	**100**	72.34	50
None	10	46.55	56.67	56.67	**100**	72.34	50
None	15	46.55	56.67	56.67	**100**	72.34	50
IEMG	–	46.55	56.67	56.67	**100**	72.34	50
SSI	–	46.55	56.67	56.67	**100**	72.34	50
Var	–	37.93	53.33	57.14	70.59	63.16	50.69
RMS	–	70.69	66.67	73.33	64.71	68.75	66.97
WL	–	89.66	**83.33**	80	94.12	**86.49**	**81.67**
LOG	–	36.21	33.33	42.86	52.94	47.37	30.32
SSC	–	74.14	80	**92.31**	70.59	80	81.45
RMS, SSC, WL, Var	–	77.59	43.33	50	11.76	48.19	78.51
IEMG, SSI, Var, RMS, WL, LOG	–	46.55	56.67	56.67	**100**	72.34	50
IEMG, SSI, Var, RMS, WL, WAMP, LOG, SSC	–	46.55	56.67	56.67	**100**	72.34	50
DWT	–	**100**	80	82.35	82.35	82.35	79.64

Table 4.4.: Test scores of the SVM model obtained for different feature sets.

4.2.5. Random Forest Classifier

Parameters

Hyperparameter Tuning with *GridSearchCV* resulted in a classifier with the following parameters. The test scores obtained for the Random Forest Classifier can be retrieved from Table 4.6.

- Classes are considered equal, class weights are balanced.
- Splitting is done using the Gini index criterion.
- Minimum 5 samples are required at the leaf node.
- Minimum 13 samples are required for splitting.
- 5 trees are contained in the forest.

Feature Importance

The impurity based feature importances extracted from the correspondant attribute of the *scikit-learn Random Forest Classifier* are illustrated in Table 4.5. The highest impact is achieved by the WL ($33.83\,\%$), followed by IEMG with $31.99\,\%$ and RMS with $21.62\,\%$. Together they make up $87.44\,\%$. The LOG detector adds $12.36\,\%$ to obtain almost $100\,\%$. The remaining features are negligable.

Feature	IEMG	SSI	Var	RMS	WL	LOG	SSC	WAMP
Feature Importance [%]	31.99	0.21	0	21.62	33.83	12.36	0	0

Table 4.5.: The impurity based feature importances

Performance

	PCA	Training Score	Test Score				
Feature(s)	**N**	**Accuracy [%]**	**Accuracy [%]**	**Precision [%]**	**Recall [%]**	**F1 [%]**	**ROC-AUC [%]**
None	–	89.66	80	86.67	76.47	81.25	80.54
None	5	82.76	76.67	72.73	**94.12**	82.05	73.98
None	10	**93.10**	83.33	80.00	94.12	86.49	81.67
None	15	84.48	63.33	66.67	70.59	68.57	62.22
IEMG	–	82.76	80	76.19	**94.12**	84.21	77.83
SSI	–	82.76	80	76.19	**94.12**	84.21	77.83
Var	–	82.76	80	76.19	**94.12**	84.21	77.83
RMS	–	82.76	80	76.19	**94.12**	84.21	77.83
WL	–	91.38	83.33	87.50	82.35	84.85	83.48
LOG	–	81.03	83.33	77.27	100	80.77	80.77
SSC	–	81.03	**86.67**	**93.33**	82.35	87.50	87.33
RMS, SSC, WL, Var	–	**93.10**	83.33	87.50	82.35	84.84	83.48
IEMG, SSI, Var, RMS, WL, LOG	–	89.66	76.67	75.00	88.24	81.08	74.89
IEMG, SSI, Var, RMS, WL, WAMP, LOG, SSC	–	**93.10**	**86.67**	84.21	**94.12**	**88.89**	**85.52**
DWT	–	86.21	76.67	91.67	64.71	75.86	78.51

Table 4.6.: Test scores of the Random Forest model obtained for different feature sets.

The best accuracy score is firstly obtained with the Random Forest Classifier is $86.67\,\%$, which was achieved once using a combination of all time-domain features, further called TDF_All, and secondly using the EMG signal frequency feature SSC. The combined feature set, however, provides a better recall score with 94.12% compared to 82.35% for SSC. Hence, its probability to correctly classify grinding and clenching is higher. The same recall score is received for IEMG, SSI, Var and RMS features alone. However, with a lower accuracy, of $80\,\%$. The utilization of PCA for dimensionality reduction of the unprocessed data to a set of 10 components, yielded a performance of $83.33\,\%$ accuracy and $94.12\,\%$ precision.

4.2.6. Summary Classification Performances

The feature sets that returned the three best test scores for each classifier are summarized in Table 4.7. Their confusion matrices are depicted in Figure 4.4-4.6.

	Highest Accuracy Score		
Classifier	$\#\mathbf{1}$	$\#\mathbf{2}$	$\#\mathbf{3}$
Logisitic Regression	$PCA_5 : 80\,\%$	$PCA_{10} : 80\,\%$	$PCA_{15} : 80\,\%$
Support Vector Machine	$WL : 83.33\,\%$	$None : 82.35\,\%$	$DWT : 82.35\,\%$
Random Forest	**TDF_All : 86.67%**	$SSC : 86.67\,\%$	$PCA_{10} : 83.33\,\%$

Table 4.7.: Summary of the best classifiers.

The table shows that the best classification performance with **Logistic Regression** was achieved using PCA. Figure 4.4 indicates that all samples for grinding and clenching were classified correctly. However, 6 samples of $Class\,1$ are falsely predicted as $Class\,0$. $53.8\,\%$. which corresponds to 7 correctly classified samples.

Principal Component Analysis also provided good results for **Random Forest**. The third best Random Forest classifier has been in fact modelled with PCA-reduced data. Out of 17 samples 16 have been correctly classified as grinding or clenching. 10 out of 13 have been rightly predicted as another movement. Although providing the same accuracy score, the second best classifier, obtained less $TP = 14$, but more $TN = 12$ (compare to Figure 4.6). Hence, SSC focusses more on the other movements and is therefore less suitable for this thesis' objective.

Figure 4.5 shows that classification with **Support Vector Machine** is best using the Waveform Length. $94.12\,\%$ of the positive samples have been correctly predicted. However, 9 out of 13 negative samples were true negatives. 4 have been falsely classified as samples of $Class\,1$. For the second best classifier using the time-frequency domain feature DWT, the following values for TP, FN, FP, and TN are obtained. $TP = 14$, $FN = 3$, $FP = 3$, and $TN = 10$, respectively. The numbers are the same for the third best model trained with unprocessed data.

The best classifier overall, which will be used for the conception of the biofeedback system prototype introduced in Section 4.4, has been achieved combining all the time-domain features presented in Section 2.2.1 and with Random Forest as classification technique.

	predicted classes	
	GC (1)	R (0)
GC (1)	17	0
R (0)	6	7

Figure 4.4: Confusion Matrices of the feature set with the highest accuracy for Logistic Regression.

	predicted classes	
	GC (1)	R (0)
GC (1)	16	1
R (0)	4	9

(a) Best performance (WL)

	predicted classes	
	GC (1)	R (0)
GC (1)	14	3
R (0)	3	10

(b) 2nd best performance (DWT)

	predicted classes	
	GC (1)	R (0)
GC (1)	14	3
R (0)	3	10

(c) 3rd best performance (None)

Figure 4.5: Confusion Matrices of the three best feature sets with the highest accuracy for SVM.

	predicted classes	
	GC (1)	R (0)
GC (1)	16	1
R (0)	3	10

(a) Best performance (TDF_All)

	predicted classes	
	GC (1)	R (0)
GC (1)	14	3
R (0)	1	12

(b) 2nd best performance (SSC)

	predicted classes	
	GC (1)	R (0)
GC (1)	16	1
R (0)	4	9

(c) 3rd best performance (PCA_10)

Figure 4.6: Confusion Matrices of the three best feature sets with the highest accuracy for Random Forest.

4.3. Experiment 3: Calibration

In the previous section different classifiers trained on data acquired from the reference subject have been compared to each other. In order to test the generalizability of the classifier graded as best, calibration was tested on two additional subjects not represented in the training data set. The results are presented below.

4.3.1. Data Set

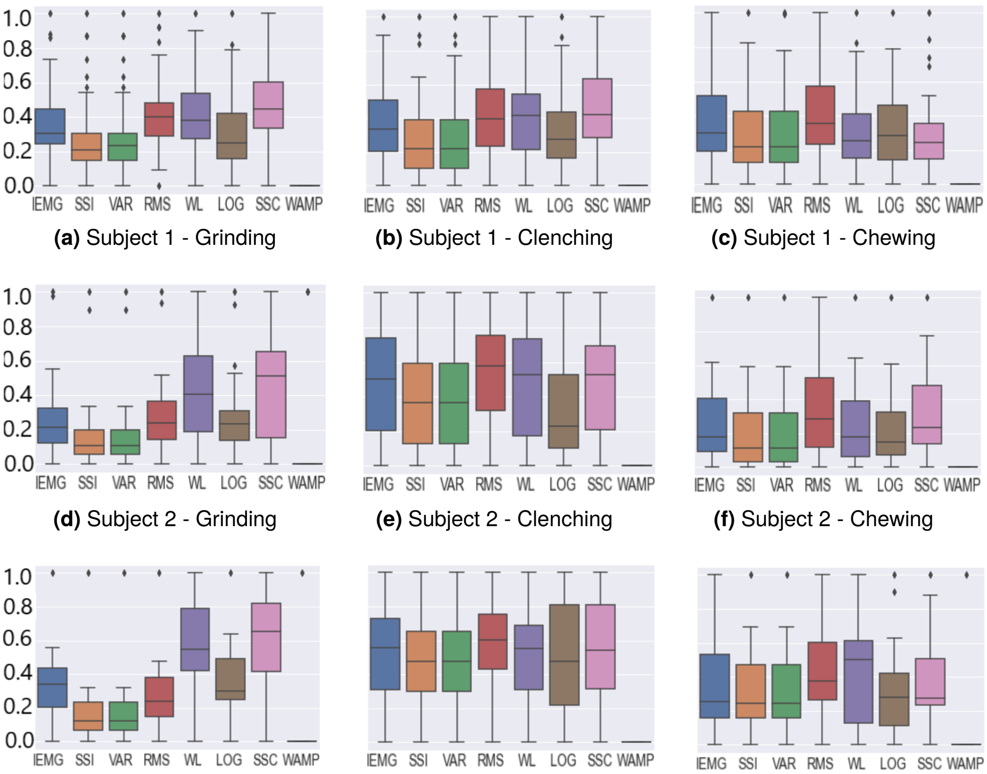
Calibration was tested on two subjects. The initial DataFrames were of shape $(54518, 5)$ and $(52921, 5)$ for Subject 1 and Subject 2, respectively. After framing 170 and 160 frames with 3144 samples were obtained. This corresponds to a $3\,s$ excerpt. Table 4.8 illustrates the structure of the data used.

	Subject 1	Subject 2
Number of Frames	170	160
Shape of DataFrame	(54518, 5)	(52921, 5)
Frame Duration [s]	3	
Number of Samples per Frame	3144	

Table 4.8.: Data Structure per Subject

4.3.2. EMG-Signals

The scaled distributions per movement of the extracted features for each subject are illustrated in Figure 4.7. Due to the lack of space, only the statistics for grinding, clenching and chewing are depicted, the distributions of the other movements are attached in Appendix A.



(a) Subject 1 - Grinding **(b)** Subject 1 - Clenching **(c)** Subject 1 - Chewing

(d) Subject 2 - Grinding **(e)** Subject 2 - Clenching **(f)** Subject 2 - Chewing

(g) Subject T - Grinding **(h)** Subject T - Clenching **(i)** Subject T - Chewing

Figure 4.7: Distribution of the extracted features for grinding, clenching and chewing determined for all subjects.

The figure shows that the lowest variance is achieved by WAMP for all movements. For grinding SSI and VAR also yield low deviations. The subplots 4.7d and 4.7g show overall a similar distribution for Subject 2 and the reference Subject T. It can be retrieved from 4.7e and 4.7h that the features have a large value range. This speaks for a heterogeneous data set. The ranges are smaller for Subject 1, which could explain the low performance of the model without calibration presented in Section 4.3.3. The subplots on chewing show that chewing differs for all the subjects. The results are similar for the attached figures.

4.3.3. Performance Results

Table 4.9 shows the results obtained when classifying the data presented in section 4.3.1 with and without calibration.

	Subject 1		Subject 2	
	Accuracy [%]	**Brier Score [%]**	**Accuracy [%]**	**Brier Score [%]**
Without calibration	60.59	37.42	83.13	11.13
Platt Scaling	61.43	**23.06**	91.67	7.10
Isotonic Regression	**70**	23.47	**93.33**	**6.08**

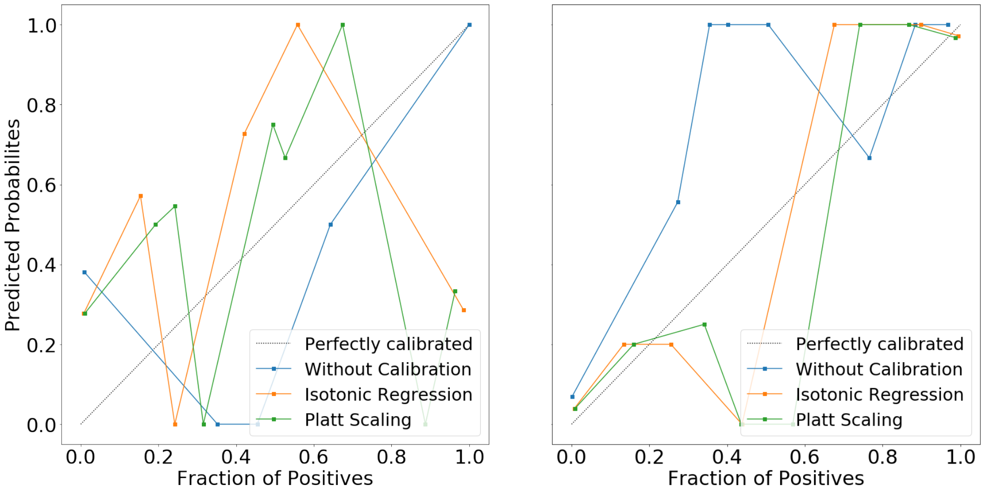
Table 4.9.: Performance of classification with and without calibration for new observations obtained from subjects, that were not represented in the training set of the classifier.

For classification the Random Forest classifier with the feature set consisting of the eight time-domain features IEMG, SSI, VAR, RMS, WL, LOG, SSC and WAMP has been chosen. The choice was made considering the classification performances presented in Section 4.2. It has been empirically shown that this combination yield best results over all combinations. The table shows that for each subject the performance of the classifier was improved using calibration. Subject 1, which showed different patterns compared to Subject 2 and Subject T (the subject on which the classifier has been trained), see Section 4.3.2 and 4.2.2, achieves overall a low performance. Without calibration, accuracy is at $accuracy_1 = 60.59\,\%$. The Brier score obtained is $BS = 37.42\,\%$, which is relatively high and can be interpreted as low performant. Platt Scaling, however, slightly increases the accuracy ($\Delta = 0.84\,\%$). The BS decreased to $23.06\,\%$, corresponding to an improvement of around $\Delta = -14.36\,\%$. The Isotonic Regression had a greater impact on the accuracy, which was increased to $70\,\%$. The improvement w.r.t. the BS was about the same magnitude as for Platt Scaling. However, it is a bit higher and thus less performant.

The uncalibrated random forest was relatively suitable for the samples recorded from Subject 2. The prediction score amounted to $accuracy = 83.13\,\%$. With $11.13\,\%$, the Brier score was also significantly lower than for Subject 1. Platt Scaling improved both accuracy ($91,67\,\%$) and BS ($7.10\,\%$). The best results were also obtained applying Isotonic Regression. An improvement of $\Delta = +10.2\,\%$ for accuracy and $\Delta = -5.05\,\%$ for BS respectively, compared to the classifier without calibration, has been achieved. The trend is, that Isotonic Regression is better suitable for the classification with a Random Forest Classifier.

Reliability of the Calibrated Classifiers

The previous results become also apparent when comparing the "reliability" curves of the calibrated classifiers, depicted in Figure 4.8.



(a) Subject 1 **(b)** Subject 2

Figure 4.8: Reliability curves of the classifier tested with data from Subject 1 and Subject 2

A binary classifier is considered as well calibrated or "reliable", if amongst the samples with a predicted probability of p, approximately $p \cdot 100\,\%$ of the samples belong to the positive class [1]. This corresponds to the linear curve in the illustration. Figure 4.8a shows that the classifier without calibration is only reliable for probabilities $p \geq 0.7$. Using Platt Scaling the predicted probabilities are closer to the fraction of positives, compared to the case without calibration, for probabilites of $0 \leq p \leq 0.6$. Isotonic Regression yields similar results. Between $0.05 \leq p \leq 0.2$ the calibration is however worse than the uncalibrated case.

From Figure 2.3b we can see, that the classifier without calibration is better calibrated for Subject 2 than the one for Subject 1. Platt Scaling and Isotonic Regression improve the reliability of the classifier, especially around $0 \leq p \leq 0.3$ for Isotonic Regression and $0 \leq p \leq 0.4$ for Isotonic Regression respectively. From $0.4 \leq p \leq 0.7$ Isotonic Regression outperforms the other cases. Starting from $p = 0.7$ Platt Scaling is again more reliable than the case without calibration. To conclude, the figure has shown, that calibration can be improved using the standard calibration techniques Platt Scaling and Isotonic Regression. However, none of them was able to calibrate the classifier such that a nearly linear curve was achieved.

[1] `https://scikit-learn.org/stable/auto_examples/calibration/plot_compare_calibration.html#sphx-glr-auto-examples-calibration-plot-compare-calibration-py`

4.4. Conception of an EMG-Biofeedback System

From above results all necessary components for the development of a biofeedback system could be retrieved. As stated in the beginning of Chapter 2, the system must contain the following elements.

1. Measurement Device

   A measurement device is needed to extract the biosignal from the subject. In this work the MyoWare Muscle Sensor (AT-04-001) has been employed. The sensor provides pre-processed data. The signal that can be retrieved from it is already amplified, rectified and integrated.

2. Data Processing Unit

   The data processing unit focuses on the transformation of the data into a suitable format for the controller. Incoming data is stored in buffers of size 3144 corresponding to signals of approx. $3\,s$ with a sampling frequency of $1048\,Hz$. The IEMG, SSI, Var, RMS, WL, LOG, SSC and the WAMP features are extracted from these buffers. This feature set is fed into the random forest model trained in the course of this work. For generalizability the classifier is calibrated using Isotonic Regression.

3. Controller

   The calibrated classifier is the core component of the controller. It returns predictions on the incoming data and sends a signal to the actuator signalling whether teeth grinding or clenching has been detected or not.

4. Actuator

   A vibration motor similar to the ones implemented in smartphones can be used for the feedback. On rising edge a pulse is send until a signal, indicating that the subject stopped clenching or grinding his or her teeth, is received.

5. Discussion

As stated in the Introduction chapter, bruxism, which is defined as excessive teeth grinding and clenching, is a sleeping disorder that can bring several health risks. In the course of this work different experiments were conducted to retrieve suitable methodologies for the detection of teeth grinding and clenching. In this chapter the obtained results presented in Chapter 4 are discussed with regard to the research questions formulated at the beginning of this thesis.

5.1. Research Question 1: Dependency between Jaw Muscle Activity and EMG Sensor Data

The first question to be discussed is whether there exists a dependency between the jaw muscle activity and the data obtained from the EMG sensor. Early experiments have shown that differences in the signal shape and amplitude can be determined visually when executing different movements. Further examinations have confirmed this hypothesis. However, the methodology applied for data acquisition presents some weaknesses.

When trying to train the classifiers, it was noticed that automatic extraction of patterns was slightly more difficult. The visual analysis of the training data set showed that the recorded signals contained longer episodes of noise, which made it harder to distinguish the movements between each other. To counteract this problem instead of automatically generating the frames and labels, framing and labelling were done manually by visual inspection. This was however only possible, because the models to be trained where supervised learning algorithms. Whether this problem also arises for unsupervised techniques was not investigated due to the small amount of data.

Although the problem only occurred for one of the three subjects, long term measures have to be considered. The data collection could be for example improved using electrode cables employed in medicine for more accurate signals. The use of shielded data cables instead of wires for data transmission could also reduce the noise level. On the software side, the recording length could be modified as follows. The time for one sample to be recorded and the pause between two recordings could be increased to $5\,s$ and $10\,s$ respectively. The samples should be saved in a DataFrame format, where the row corresponds to one signal sample or frame. Each movement should be stored in an individual DataFrame. An automatic script could then modify the samples cutting out the noise, taking into consideration that frame size should be 3144 for each frame. These measures are for instance not only important when collecting data for a larger study, but also for the collection of the calibration data set.

To sum up, one can conclude that there is in fact a dependency between the activity of the temporalis muscle, which is the jaw muscle investigated in this thesis, and the amplitude values received from the EMG sensor data. Different patterns were detected for the varying movements. Hence, Surface Electromyography is a suitable technique for the measurement of muscle activity, with the objective of detection of teeth grinding and clenching. However, it should be taken into account that results are hypothetically less accurate than with invasive EMG.

5.2. Research Question 2: Comparison of Machine Learning Techniques

The second research question deals with the determination of the best suitable machine learning method for the detection of teeth grinding and clenching. Relying on applied algorithms for the detection of other disorders in medicine, three methods were chosen for evaluation of their applicability for the detection of teeth grinding and clenching. Several papers have shown that Support Vector Machines(SVM) are a great basis to start from. SVM is a classification technique that aims to maximize the margin between the class borders and the samples closest to it. The method accepts linear and non-linear borders. Non-linear SVM are for example the Gaussian radial basis function (RBF), the sigmoid function or polynomial functions. They differ in the shape of their class borders. Parameter Tuning has shown that a polynomial function of $4th$ degree was best suited for the discrimination of both classes (1: teeth grinding and clenching, 0: other movements). The training of a SVM model however is very cost intensive, especially with a polynomial of 4 and yields lower performance than its contractor Random Forest.

Random Forest is a classification technique based on independent and random decision trees. The risk of overfitting is reduced by the fact that new classes are predicted following a majority vote of the whole forest and is minimal in case of independent trees, which is probably the reason why this technique worked best for the thesis' use case. It has been found that the classifiers are most performant when using feature extraction. Hence a model with a $(1, 8)$ feature set consisting of IEMG, SSI, Var, RMS, WL, LOG, SSC and WAMP has been used. The time-domain features provided better results than its time-frequency opponent DWT. The utilization of the uncompressed signals of size $(1, 3144)$ yielded acceptable results for SVM and Random Forest. They were however outperformed by the model with the reduced feature set.

Surprisingly, the third classification technique Logistic Regression, that was chosen to evaluate the performance of a linear classification method, has shown its best result using PCA for dimensionality reduction. The accuracy changed from $53.33\,\%$ to $80\,\%$. Nevertheless, the performance was lower than Random Forest.

To conclude, differences between the performances of the classifier were identified. A model achieving $86.67\,\%$ of accuracy and $94.12\,\%$ of recall is obtained. However, it is important to take into consideration that the manual labelling of the training set could bring biased results. The small size of the training data set might also affect the results.

5.3. Research Question 3: Impact of Classifier Calibration

The last research question covers the impact of model calibration with regard to its generalizability. The aim was to check, whether calibration was necessary regarding the fact that above mentioned algorithms were trained using one subject only and if it could improve performance. The study has shown that calibration of classifiers is indeed necessary. This is mainly due to variants of the EMG signal for each person and circumstance. Although the trained model was also relatively well suitable for one subject $accuracy = 83.13\,\%$ and $BS = 11.13\,\%$, it performed poorly on the other ($accuracy = 60.59\,\%$ $BS = 37.42\,\%$).

For calibration the standard methods Platt Scaling and Isotonic Regression were tested. The former, which is also known as sigmoid scaling, provided improvements for both subjects. The BS of the more diverse subject decreased from $37.42\,\%$ to $23.06\,\%$ which corresponds to a better calibration. However, the accuracy did not show much improvement ($\Delta < 1$). Better results were obtained using Isotonic Regression. With this technique the accuracy in the data set for Subject 1 increased from $60.59\,\%$ to $70\,\%$. The calibration score showed also an improvement compared to the case without calibration. The accuracy on the data of Subject 2 could even be increased to $93.33\,\%$ without neglecting the BS which reached amongst all options the lowest value of $6.08\,\%$.

The trend is that Isotonic Regression works best for the calibration of a random forest for teeth grinding and clenching detection. Compared to Platt Scaling it does not only focus on the enhancement of the calibration as measured by the Brier score, but also significantly improves the classification accuracy. This trend could however be misleading as calibration was only tested on 2 subjects. Hence, it should be evaluated on more persons. The results obtained from the reliability curves show that none of the techniques was able to achieve a nearly linear curve and thus a well calibrated model. Hence, calibration might be enhanced using other methods. The applicability of the calibration methods of Dankowski and Ziegler [4] and/or Boström [1], presented in Section 2.7.3, might be interesting to investigate in the future.

6. Conclusion and Future Work

The objective of this thesis was the examination of possible techniques for the detection of bruxism using sEMG. Regarding the fact, that the term defines **excessive** teeth grinding and clenching, the focus was set on the detection of the common execution of these movements, to avoid health consequences.

The study has been conducted using the MyoWare Muscle Sensor (AT-04-001) with three silver-chloride electrodes. Experiments have shown that better signals were obtained when placing two electrodes on the temporalis muscle, situated on the temple, and the third one, the reference electrode, on the forehead. Although, visual inspection has revealed that all movements apart from resting showed similarities, slight differences in amplitude, period and pulse width could be found. A cluster analysis of the PCA reduced data with DBSCAN however showed that the two classes *bruxism* and *no bruxism* are distinguishable.

In order to evaluate, which machine learning algorithm is best suitable for the detection of teeth grinding and clenching, Logistic Regression, Support Vector Machine and Random Forest classifiers with different inputs were trained and tested. The data has been acquired from one subject on different days and at different times. Significant signals of the recordings have been manually detected and labelled. Varying patterns for each movement have been recognized visually. Statistical confirmation has been obtained analyzing the signal amplitudes. The highest amplitudes were found in yawning, grinding, chewing and clenching, in descending order. The average maximum peak for each movement showed that yawning, clenching and grinding showed the highest amplitudes.

Logistic Regression achieved best performance in terms of accuracy using PCA-reduced data. The dimensionality has been reduced to 5, 10 and15 principal components, all yielding an accuracy of $80\,\%$. With SVM slightly better results were obtained. The highest score with $83.33\,\%$ was achieved using the accumulated waveform length above a time segment (WL) as input. Even better results were perceived with the Random Forest algorithm. Most of the tested inputs provided accuracy values higher than $75\,\%$. The best score of $86.67\,\%$ was obtained using a 1×8 feature set composed of the time domain features IEMG, SSI, Var, RMS, WL, WAMP, LOG and SSC.

The generalizability of the Random Forest model graded as best was evaluated using data acquired from two further subjects, that was not represented in the training set of the classifier. Without calibration the model was able to classify the samples from the first subject with an accuracy of $60.59\,\%$. For the second subject the value was at $83.13\,\%$. Calibration using Platt Scaling improved the Brier score. It has nevertheless been outperformed by Isotonic Regression, which additionally optimized the accuracy. However, calibration might be enhanced using other calibration techniques specialized in the calibration of random forests. To sum up, the detection of teeth grinding and clenching using surface EMG works best using a calibrated Random Forest classifier with 5 trees and the Gini index as splitting criterion. Isotonic Regression showed promising results. However, it has to be evaluated on further subjects to get solid results.

This thesis has shown that a relation between jaw movement and EMG data exists and that this relation can be used to detect teeth grinding and clenching. Owing to the the inability to conduct a study due to the pandemic an univariate classification problem was faced. A study with a larger set of features (multivariate classification problem) could be of interest to find additional information and relations between teeth grinding and clenching and other variables such as gender, age or even a subject's health condition. Moreover, the study should aim to have a larger number of subjects. The subjects should be chosen such that there can be a variety of study variables. To maximize the variance the subjects should be recorded twice with a certain time difference. This may lead to a more generalized detector and to more accurate results. Another point of interest is the examination of the effect of biofeedback. This will require to develop the concepted biofeedback system into a wearable device, which will ease the EMG recording at home. The data acquired by this can additionally be used to correct and update previous detectors. The study should be conducted in close cooperation with specialists in the field of sleep bruxism to assure that the results are in coherence with their knowledge.

Bibliography

[1] H. Boström. “Calibrating Random Forests”. In: *Seventh International Conference on Machine Learning and Applications, 2008*. Ed. by M. A. Wani. Piscataway, NJ: IEEE, 2008, pp. 121–126. ISBN: 978-0-7695-3495-4.

[2] T. Castroflorio, L. Mesin, G. M. Tartaglia, C. Sforza, and D. Farina. “Use of electromyographic and electrocardiographic signals to detect sleep bruxism episodes in a natural environment”. In: *IEEE journal of biomedical and health informatics* 17(6) (2013), pp. 994–1001.

[3] R. H. Chowdhury, M. B. I. Reaz, Ali, Mohd Alauddin Bin Mohd, A. A. A. Bakar, K. Chellappan, and T. G. Chang. “Surface Electromyography Signal Processing and Classification Techniques”. In: *Sensors* 13(9) (2013), pp. 12431–12466.

[4] Dankowski, T. and Ziegler, A. “Calibrating random forests for probability estimation”. In: *Statistics in Medicine* 35(22) (2016), pp. 3949–3960. ISSN: 1097-0258.

[5] Dorr, M. *Pattern Recognition: Lecture Notes*. München, 2018.

[6] P. G. Fonseca and H. D. Lopes. *Calibration of Machine Learning Classifiers for Probability of Default Modelling*. 2017.

[7] D. G. Kleinbaum, M. Klein, and E. R. Pryor. *Logistic Regression: A Self-Learning Text*. 3. ed. Statistics for Biology and Health. New York, NY: Springer Science+Business Media LLC, 2010. ISBN: 9781441917416.

[8] Konrad, P. *The ABC of EMG: A Practical Introduction to Kinesiological Electromyography*. 1.4. USA: Noraxon INC. USA, 2006. ISBN: ISBN 0-9771622-1-4.

[9] C. L. Long. “Pattern recognition using surface electromyography of the anterior temporalis and masseter muscles”. Ph.D. thesis. 2004.

[10] D. Manfredini, E. Winocur, L. Guarda-Nardini, D. Paesani, and F. Lobbezoo. “Epidemiology of bruxism in adults: a systematic review of the literature”. In: *Journal of orofacial pain* 27(2) (2013), pp. 99–110. ISSN: 1064-6655.

[11] Momen, Kaveh Syeh and Chau, Tom Tak Kin. “Method, system and apparatus for real-time classification of muscle signals from self-selected intentional movements”. (US8437844B2). 2007.

[12] N. Nazmi, M. Abdul Rahman, S.-I. Yamamoto, S. Ahmad, H. Zamzuri, and S. Mazlan. “A Review of Classification Techniques of EMG Signals during Isotonic and Isometric Contractions”. In: 16 (Aug. 2016), p. 1304. ISSN: 1424-8220.

[13] S. E. Nishi, R. Basri, and M. K. Alam. “Uses of electromyography in dentistry: An overview with meta-analysis”. In: *European journal of dentistry* 10(3) (2016), pp. 419–425. ISSN: 1305-7456.

[14] Ober, Stephen H. “Apparatus and method for preventing bruxism”. (US4669477A). 1987.

[15] M. M. Ohayon, K. K. Li, and C. Guilleminault. “Risk factors for sleep bruxism in the general population”. In: *Chest* 119(1) (2001), pp. 53–61. ISSN: 0012-3692.

[16] S. Orguc, H. S. Khurana, K. M. Stankovic, H. S. Leel, and A. P. Chandrakasan. “EMG-based Real Time Facial Gesture Recognition for Stress Monitoring”. In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. 2018, pp. 2651–2654.

[17] Peroz, I., Bernhardt, O., Kares, H., Korn, H. J., Kropp, Lange, M., Müller, A., Nilges, P., Ommerborn, M. A., Steffen, A., Tholen, R., Türp, J. C., and Wolowski, A. *Diagnostik und Behandlung von Bruxismus*. 2019.

[18] A. Phinyomark, P. Phukpattaranont, and C. Limsakul. “Feature reduction and selection for EMG signal classification”. In: *Expert Systems with Applications* 39(8) (2012), pp. 7420–7431. ISSN: 09574174.

[19] A. Phinyomark, F. Quaine, S. Charbonnier, C. Servière, F. Tarpin-Bernard, and Y. Laurillau. “EMG feature evaluation for improving myoelectric pattern recognition robustness”. In: *Expert Syst. Appl.* 40 (2013), pp. 4832–4840.

[20] E. Scheme and K. Englehart. “Electromyogram pattern recognition for control of powered upper-limb prostheses: state of the art and challenges for clinical use”. In: *Journal of rehabilitation research and development* 48(6) (2011), pp. 643–659.

[21] H. Sharma and S. Kumar. “A survey on decision tree algorithms of classification in data mining”. In: *International Journal of Science and Research (IJSR)* 5(4) (2016), pp. 2094–2097. ISSN: 2319-7064.

[22] R. E. Singh, K. Iqbal, G. White, and J. K. Holtz. *A Review of EMG Techniques for Detection of Gait Disorders*. 2019.

[23] K. Stapor. “Evaluation of classifiers: current methods and future research directions”. In: Sept. 2017, pp. 37–40.

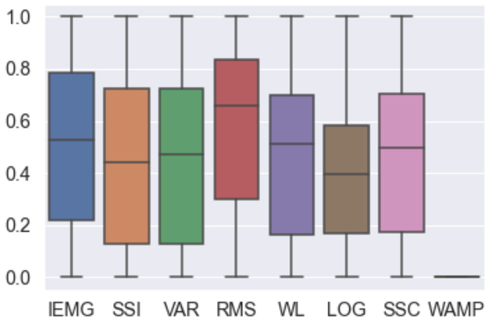
[24] Thumedi & Co KG GmbH. “Measuring, analysis and biofeedback device”. (DE202009003444U1). 2009.

[25] D. C. Toledo-Pérez, J. Rodríguez-Reséndiz, R. A. Gómez-Loenzo, and J. C. Jauregui-Correa. “Support Vector Machine-Based EMG Signal Classification Techniques: A Review”. In: *Applied Sciences* 9(20) (2019), p. 4402.

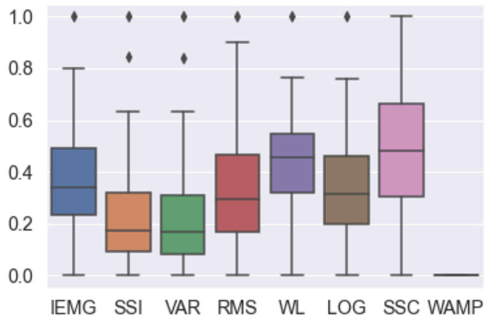
[26] Wong. “Electromechanical bruxism monitoring and treating equipment”. (CN1810319A). 2005.

A. Scaled Distributions per Movement

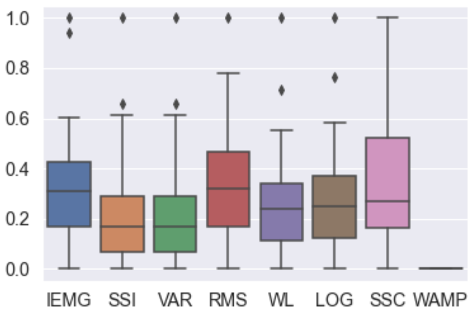
Depicted are the scaled distributions for the remaining movements (yawning and resting) for each subject.



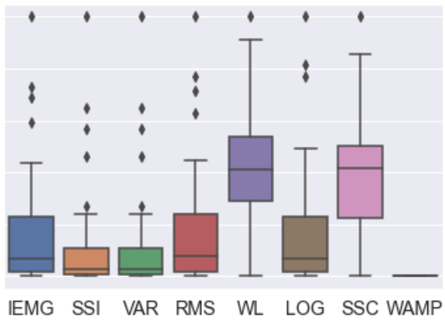
(a) Subject 1 - Yawning



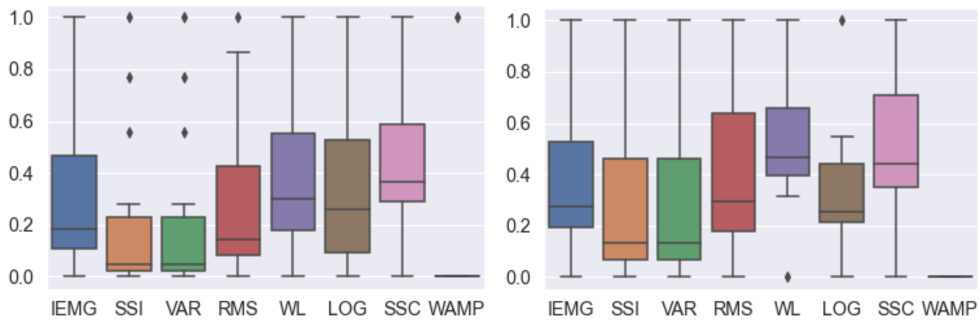
(b) Subject 1 - Resting



(c) Subject 2 - Yawning



(d) Subject 2 - Resting



(e) Subject T - Yawning

(f) Subject T - Resting

Figure A.0: Distribution of the extracted features for yawning and resting determined for all subjects.