Shobiha K. Premkumar

EEG-based Biometric System for Subject Identification using Empirical Mode Decomposition and Frequency Bands

Specialization project

Supervisor: Marta Molinas
Co-supervisor: Luis Alfredo Moctezuma

Department of Engineering Cybernetics
Norwegian University of Science and Technology

Abstract

This work investigates the use of brain activity signals as a suitable parameter for a biometric system. Brain activity captured in electroencephalography (EEG) data can be analyzed using feature extraction and classification to identify a subject.

This work presents signal analysis methods and features for analyzing EEG signals on two different neuro-paradigms; resting-state and event-related potential. The EEG signals are decomposed using the Empirical Mode Decomposition (EMD) and frequency bands. Features (energy, fractal, statistical and HHT-based) are then extracted from the decomposed signals and are used as input on five different for machine learning algorithms (DT, RF, SVM, k-NN, and NB) for obtaining trained models. Machine learning is utilized to identify the unique patterns in EEG-signals. The model with the highest accuracy is utilized for validation with unseen data.

Using frequency bands and EMD as the basis for feature extraction on resting-state data, the highest validation accuracy obtained was 0.98 and 0.92, respectively, and 0.95 and 0.89 for the event-related potential. The findings in this project reveal important information for continuing the exploration of feature extraction for Subject Identification.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The ability to identify individuals is of the highest importance in both government and civilian applications. Traditionally, a combination of knowledge-based methods (e.g., PINs and passwords) and token-based methods (e.g., keys and ID-cards) has been used to validate the identity of an individual. In a broad-scale application, like border control, where thousands of people are inspected, the traditional methods are vulnerable for imposters and spoofing. Replacing conventional authentication methods with biometrics is therefore introduced [1, 2, 3].

Biometrics refers to identification based on certain physical or behavioral traits of an individual. By using biometrics, it is possible to authenticate the individual's identity based on "who you are" rather than "what you possess" (e.g., ID card) or "what you remember" (e.g., password) [4]. Currently, biometrics traits such as fingerprints, facial features, voice, and DNA, are adopted in real-life scenarios [5]. However, even biometrics are prone to spoofing and can easily be stolen, similar to traditional methods. Once these biometrics are stolen, they can not be replaced like conventional tokens; an ID-card can be replaced in contrast to growing a new fingerprint pattern.

Any physical or behavioral traits can be used as a biometric as long as some criteria are met: it is challenging to steal, and it is cancelable. A biometric trait fulfilling these criteria are brain signals, which can be measured from the scalp using a technique known as Electroencephalography (EEG).

Brain signals obtained using EEG appears from brain activities created from unique patterns of neural pathways. The resulting brain activities are unique for every person and are related to the subjects' genetic information. Because of its uniqueness, brain signals makes a good base as a biological feature for subject identification [6]. There are several advantages for using brain activity as biometric measurement compared to biometrics used today:

1. A person has to be alive to be able to produce EEG signals; lack of EEG is an indication of brain death. This protects the user from being dead and unconscious to provide valid EEG data.

2. The brain activity is measured as a voltage. Increasing the distance from the scalp will decrease the measurable voltage. The EEG needs contact for collecting data.

3. Brain signals can be elicited by numerous separate brain systems, which makes brain signals cancelable.

4. EEG signals are sensitive to stress, which protects the user from being forced.

A biometric system consists of two parts: the data acquisition part and the decision part. Data acquisition consists of recording brain activity while a subject engages with a protocol, such as resting-state, motor imagery, or visual stimulation. The decision part is where the acquired data is pre-processed for increasing the Signal-to-Noise ratio (SNR), as recorded EEG signals are prone to noise. The decision part also consists of feature extraction to obtain characteristics of the unique EEG signals. The different sets of features are then categorized by a model created using machine-learning techniques. The trained model is used to identify a subject by entering new input data.

The utilization of brain activity as biometrics with different neuro-paradigms (e.g., resting-state, imagined speech, color exposure) is of significant interest. In the time being, the utilization of brain biometrics is still not possible in real life. Brain biometrics is still placed in the research field as it has multiple factors to improve. Prior research has been able to demonstrate uniqueness, permanence, and universality of using brain signals. For the possibility to apply brain biometrics in real life, the collectability, performance, and acceptability of brain signals have to be improved through research.

1.1 Problem description

The purpose of this project is to investigate feature extraction methods and classification algorithms on different neuro-paradigms using EEG signals.

The aim of this project is subject identification using EEG signals from different neuro-paradigms. The problem is approached by using signal analysis methods for getting meaningful physical signals from EEG signals and then extract features for classification. A variety of features and classification algorithms are explored on two datasets containing two types of neuro-paradigms: resting-state and cognitive task.

The evolution of classification accuracy when the number of EEG recordings channels are reduced is also investigated.

1.1.1 Motivation

The methods proposed in section 1.1 is an approach for real-time subject identification using EEG signals with a reduced number of EEG recordings channels. For real-time identification, the device used for recording EEG signals should be portable with few channels, and the response time from the system classification should short. Using relevant features is one way for reducing computation time and obtaining a better representation of the EEG signals. A better representation of EEG signals can provide higher identification accuracy.

1.1.2 Report structure

This report starts with relevant background and theory about biometric systems, EEG signals, as well as signal analysis methods and machine learning algorithms in Chapter 2. Relevant work in subject identification using EEG signals are presented in Chapter 3. In Chapter 4, the materials and methods used in this project are described. The obtained results are presented in Chapter 5, with a discussion of the results in Chapter 6. Chapter 7 concludes this project and presents future work.

Chapter 2

Background and theory

2.1 Biometrics

Biometrics is the technical term for the identification of individuals based on their physiological or behavioral characteristic. [4]. Any human physiological or behavioral characteristic can be used as a biological measurement as long as the following requirements, as mentioned in [4] and [1] are satisfied:

- *Universality*: the characteristic should exist in every individual.
- *Uniqueness*: no other individuals can be equal in terms of the characteristic.
- *Permanence*: the characteristic should be invariant (to the matching criterion) over some time.
- *Collectability*: the characteristic can be measured quantitatively.

In terms of a practical biometric system, other essential requirements should be considered as well, such as:

- *Performance*: the achievable identification accuracy and speed, the requirement for recourses to achieve an acceptable accuracy and speed, and working or environmental factors that affect the identification accuracy and speed.
- *Acceptability*: to what extent are people willing to accept the use of particular biometric characteristic.
- *Circumvention*: how easily the system can be fooled by spoofing.

2.2 Biometric system

A biometric system employs biometrics for subject identification. This system may be referred to as a pattern recognition system whose function is to classify a biometric signal into several identities [1]. A biometric system is designed consisting of four central modules which are presented in Figure 2.1:

1. A sensor module capturing the raw biometric data from a subject.

2. A feature extraction module that extracts a set of features representing the acquired biometric signal. The extracted features are labeled with the identity of the subject and stored in the biometric system as a template.

3. A matching module generating matching scores by comparing the extracted features from authentication with the stored templates.

4. A decision module processing the calculated matching scores in order to verify or determine the identity of the subject.

The system can operate as verification or identification depending on the context of the application:

- *Validation mode*: the individuals' identity (e.g., ID-card or Personal Identification Number (PIN)) is validated by comparing the captured biometric data with the individuals' biometric templates stored in the database.

- *Identification mode*: the system will search through all the stored templates in the database to find a match for recognizing an individual.

In this project, a biometric system based on identification mode is chosen. Without the subject having to claim an identity, the system carries out a one-to-many comparison to establish the subject's identity. Individuals not enrolled in the system will then fail. Identification is known as negative recognition to prevent a single individual from using multiple identities [7]. Negative recognition can only be performed with biometrics.

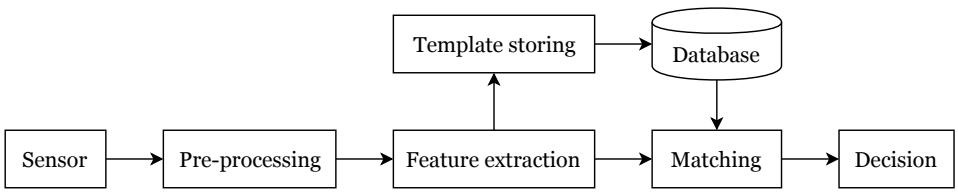

Figure 2.1: Main modules of a biometric system

2.3 Electroencephalography

An electroencephalogram is a technique for measuring the electrical activity generated by the brain. This method is non-invasive by placing electrodes on the scalp for recording the EEG signals [8]. The electrode placement on the scalp follows the international 10-20 system standardized by the American Electroencephalographic Society [9], as shown in Figure 2.2.

It is challenging to gain useful information from EEG-signals directly in the time domain just by observation. Raw EEG signals are both nonlinear and non-stationary by nature with a small amplitude since the signals have to cross scalp, skull, in addition to many other layers. EEG signals are, therefore, prone to background noise and artifacts occurring both internally and externally [10]. Artifacts contaminating the EEG signals could be muscle movement, blinking, and face movements and external noise could be the electrical noise from powerline at 50 Hz or 60 Hz [11].

The acquired EEG signals are different for different brain activities. One way to study the brain is by triggering different simulations by presenting a paradigm, such as motor imagery, event-related response, and visual evoked potentials.

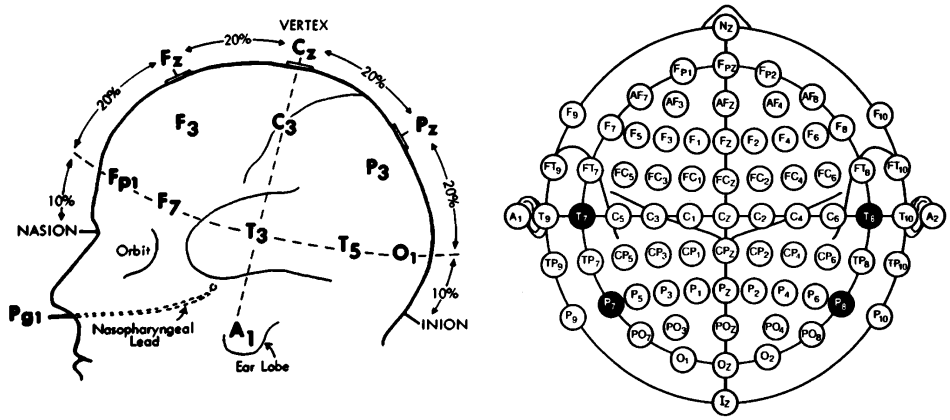

Figure 2.2: Electrode placement on the scalp according to the international 10-20 system. Left image lateral view, right image top view [12]

2.3.1 Frequency bands of the brain

Raw and unprocessed EEG signals from the brain can reveal neural oscillations, which is always a mixture of underlying base frequencies. These frequency bases reflect the different states of our brain that varies depending on individual factors, stimulus properties, and internal states. Brain waves are, therefore, characterized into the following frequency bands sorted from lower to higher frequencies: delta, theta, alpha, beta, and gamma bands. These are referred to as the frequency bands

Brain rhythms	Frequency [Hz]	Description
Delta	0.5 - 4.0	Deep sleep
Theta	4.0 - 8.0	Memory demands
Alpha	8.0 - 12.0	Awake, relaxed
Beta	12.0 - 30.0	Alertness and focused attention
Gamma	>30.0	Deep focus

Table 2.1: Brain frequency bands and their respective frequency range.

of the brain. Depending on what brain activity is induced, different frequency bands to the specific cognitive process will be active [13]. Each frequency band is correlated with their associated mental state presented in Table 2.1 [5].

2.3.2 Event-related potential

An event-related potential (ERP) is a time-locked EEG signal which captures neural activity related to specific events or stimuli. They are of small voltages and are utilized for the evaluation of brain functions and response to stimuli. The presented stimuli generate detectable but time-delayed waves in EEG signals and indicate how the stimulus is processed. A well known ERP wave pattern is the P300 peak. The P300 component occurs approximately 300 ms after a stimulus is delivered, and appears as a series of positive and negative voltage fluctuations in the EEG signal [14].

2.4 Signal Analysis Methods

As mentioned in section 2.3, EEG signals do not provide any useful information in their original form because of their natural shape and added noise. To be able to classify subjects using EEG signals, features extraction with advanced signal processing techniques as a basis are required [11]. The results obtained from signal analyzing depends on the applied signal analysis method, the experiment, and the signal characteristics. When analyzing EEG signals, both high-frequency resolution and high time resolution is of interest.

2.4.1 Fast Fourier Transform

The Fourier transform (FT) transforms a function of time from the time domain into the frequency domain. The hidden information in the frequency domain can then be extracted and analyzed. The Discrete Fourier Transform (DFT) is used when dealing with a finite sequence of equally-spaced samples signals with the formula given in Equation 2.1:

$$X_k = \sum_{n=0}^{N-1} x_n \cdot e^{-\frac{i2\pi}{N}kn} \tag{2.1}$$

where N is the number of complex number $x_n := x_0, x_1 ... x_{N-1}$ transformed into an another sequence of complex number $X_n := X_0, X_1 ... X_{N-1}$. The computational cost of DFT is $\mathcal{O}(N^2)$, where N is the data size. The Fast Fourier Transform (FFT) algorithm is therefore used for computing the DFT, where it computes all DFT coefficient as a "block" with a computational cost proportional to $\mathcal{O}(Nlog_2N)$ [15].

The FFT only provides information limited to the frequency domain. For analyzing EEG signals, both time and frequency domain is needed to extract useful information about the signal. FFT is a useful method for examining the content of different frequency components in EEG signal.

2.4.2 Empirical Mode Decomposition

The Empirical Mode Decomposition is an adaptive method for decomposing nonlinear and non-stationary time-series data, such as EEG signals. The data is decomposed into several Intrinsic Mode Functions (IMFs) without leaving the time domain, and must satisfy two conditions [16]:

- Condition 1: The number of local minima and maxima differs at most by one.
- Condition 2: At any point, the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero

Decomposing signals into IMFs makes EMD a data-driven method that does not depend on any *a priori* defined basis system. The IMFs are extracted through a process called *Sifting*, which removes riding waves and make the wave-profile more symmetric [17] [16]. The sifting process outputs IMFs through an iterative procedure and works as follows:

1. Identify all the local extrema in the signal
2. Compute lower and upper envelopes from interpolations between extrema; $e_{lower}(t), e_{upper}(t)$
3. Calculate the local mean value with the lower and upper envelope; $m_{1,1}(t) = 0.5(e_{lower}(t) + e_{upper}(t))$
4. Subtract the mean value from the signal; $h_{1,1}(t) = x(t) - m_{1,1}(t)$
5. Determine if the extracted signal is an IMF with the given conditions of an IMF (Condition 1 and condition 2)
6. Repeat step 1 - 4 until an IMF is obtained; $c_1(t) = h_{1,k}(t)$
7. Subtract the obtained IMF from the original signal; $x_2(t) = x(t) - c_1(t)$
8. Repeat steps 1- 6 until there are no more IMFs to extract. The last component extracted as an IMF is called residual.

When decomposition of n IMFs are finished, the original signal can be reconstructed as

$$x(t) = \sum_{i=1}^{n} c_i(t) + r_n(t) \tag{2.2}$$

Limitations with EMD

The spline interpolation in the sifting process is an approximation, which leads to some minor deviation from the real mean envelope. End effects are a difficulty with EMD that occurs near the ends of the signal, which can make the spline interpolation produce large swings. A solution for end effects is presented in [16].

Another difficulty with EMD is the mode mixing problem during the sifting process. The mode mixing problem occurs when the data contains intermittency and can make the IMF lose physical meaning. Data affected by noise can also cause mode mixing, as it can be thought of as another kind of intermittency. A method for removing mode mixing is proposed in [18].

A more robust method, which is a further development of EMD is the ensemble mode decomposition (EEMD). The EEMD is utilized for removing noise and mode mixing and defines true IMFs components as a mean of an ensemble of trials [19]. A random white noise of finite-amplitude is added to the signal in each trial, and EMD is then applied to this signal. When all the trials are finished, an overall mean is then calculated for obtaining the true result. However, the computation of EEMD is more complex than EMD because of the ensemble number of trials and therefore not suitable for real-time application [20]

2.4.3 Hilbert-Huang Transform

Instantaneous frequency

A proper definition of instantaneous frequency is of interest when analyzing nonlinear dynamical systems like brain activity. The recorded EEG signals contain multiple frequencies that could exist at the same time, and instantaneous frequency is therefore necessary. One method for obtaining this is using the Hilbert Transform.

Hilbert Transform

Hilbert Transform (HT) can be applied to a signal for generating an analytic signal [17]. The analytic signal $z(t)$ is obtained by adding the original signal $x(t)$ with the imaginary part of the transformed signal $y(t) = \mathcal{H}\{x(t)\}$ as shown in Equation. 2.3:

$$z(t) = x(t) + i \cdot y(t) = a(t)e^{i\phi(t)} \tag{2.3}$$

where

$$a(t) = \sqrt{x^2(t) + y^2(t)} \tag{2.4}$$

$$\phi(t) = \arctan\left(\frac{y(t)}{x(t)}\right) \tag{2.5}$$

$$\omega(t) = \frac{d\phi(t)}{dt} \tag{2.6}$$

represents the instantaneous amplitude, the instantaneous phase, and the instantaneous frequency of the signal, respectively. The aim is to obtain meaningful instantaneous frequencies that are local [16].

Hilbert-Huang transform

IMF obtained from EMD, represents one of the oscillatory modes in a nonlinear and non-stationary signal. These can be both amplitude and frequency modulated. The IMFs do not provide any good physical interpretation of the data on their own and need to be further analyzed.

Taking the HT of a real-valued signal like IMF, the obtained analytic signal can then be used to extract the instantaneous frequency as a function of time. Since the extracted IMFs are obtained from local properties, the instantaneous frequency of the signal will provide meaningful information about the complicated signal. Any event can be localized in time, as well as the frequency axis. This combination of using IMFs from EMD and the HT is known as the *Hilbert-Huang Transform* (HHT) [16].

2.5 Features extraction

EEG signals are further analyzed using feature extraction. A feature represents an individual measurable property of a process being observed [21]. Recorded EEG signals contain several different features which can be used for representing the signals. Machine learning algorithms can perform classification on EEG signals by using a set of features. Searching for a limited amount of features representing the signal with certainty is necessary for reducing computations. The process for selecting relevant features are called feature selection and helps to understand the data, reduces the computational requirement, removes irrelevant or redundant variables, and improves the predictor performance [21].

2.5.1 Energy features

Energy features are used for extracting the amplitude and frequency information from EEG signals. The *instantaneous energy* gives information about the signal amplitude and is computed as shown in Equations 2.7

$$f = log_{10}\left(\frac{1}{N}\sum_{i=1}^{N}(vec(i))^2\right) \tag{2.7}$$

where N is the length of a vector and vec is the coefficient of a vector at position i [22]. The *Teager energy* describes variations in signal frequency and is defined as

$$f = log_{10}\left(\frac{1}{N}\sum_{i=1}^{N-1}\left|(vector(i))^2 - vector(i-1)\cdot vector(i+1)\right|\right) \tag{2.8}$$

2.5.2 Fractal features

The fractal dimension describes how a measure of a time series, such as EEG signals, changes depending on a scale used as a unit of measure, described in the form of a complex index [23]. The *Petrosian fractal dimensions* (PFD) and *Higuchi fractal dimensions* (HFD) are two types of fractal dimensions used in this project.

The PFD provides a fast computation of the fractal dimension of a signal by translating the signal into a binary sequence. The binary sequence is created by assigning '1' when the difference between sequential samples in the signal exceeds a standard deviation magnitude, and a '0' otherwise [24]. The PFD is computed as shown in Equation 2.9

$$FD_{Petrosian} = \frac{log_{10}n}{log_{10}n + log_{10}\left(\frac{n}{n+0.4N\Delta}\right)} \tag{2.9}$$

where n is the length of the sequence and $N\Delta$ is the number of sign changes in the binary sequence.

The HFD approximates the mean length of a signal using segments of k samples and estimates the dimension of a time-varying signal directly in the time domain, which reduces the running time [25] [26]. The N-sampled data sequence $X(1), X(2), ..., X(N)$ is divided into new time series that are subsets of k samples and are constructed as follows:

$$X_k^m : X(m), X(m+k), X(m+2k), ..., \left(X\left(m + \frac{N-m}{k}\right)k\right) \tag{2.10}$$

where $m = 1, 2, ..., k$ is the initial time and $k = 1, ..., k_{max}$ is the interval time with k_{max} being a constant parameter. In this project $k_{max} = 10$ was used. The length

$L_m(k)$ is then calculated for each subset X_k^m as:

$$L_m(k) = \frac{1}{k}\left(\sum_{i=1}^{\frac{N-m}{k}} |x(m+ik) - x(m+(i-1)k)|\right)\left(\frac{N-1}{\frac{N-m}{k}}\right) \tag{2.11}$$

The mean value array for the overall signal is then calculated:

$$L_k = \frac{1}{k}\sum_{m=1}^{1} L_m(k) \tag{2.12}$$

The HFD is estimated using the array of mean values L_k by calculating the least square slope of the trajectory:

$$FD_{Higuchi} = \frac{ln(L_k)}{ln(\frac{1}{k})} \tag{2.13}$$

2.5.3 Statistical features

Statistical measurement can be used to extract different features from EEG signals. The statistical features used in this work are presented in Table 2.2.

Features	Description
Maximum, minimum	Highest and lowest potential in a time series.
Mean, median	Central tendency, middle score for a set of a data arranged in order of magnitude.
Variance, standard deviation	Dispersion around the mean.
Kurtosis	Measure of whether the data is light-tailed or heavy-tailed relative to a normal distribution.
Skewness	Measure of lack of symmetry.

Table 2.2: Statistical features used in this project.

2.5.4 HHT-based features

Two features are computed based on HHT. The *marginal frequency* is obtained by computing the sum of the instantaneous frequencies from each IMF. The *mean instantaneous amplitude* is computed for each IMF. These features are recreated from [27].

2.6 Machine learning

According to [28], machine learning is a computers' ability to adapt to new circumstances and to detect and extrapolate patterns. By exposing a subject to different

paradigms, EEG signals can be recorded containing different patterns and use machine learning to reveal these patterns. Machine learning gives computers the ability to learn from experience by using one of two types of learning techniques:

- *Supervised learning*: known input and output data are used for training a model for predicting future outputs.
- *Unsupervised learning*: hidden patterns are detected from input data.

Extracted feature vectors with corresponding target label are used as training parameters for a model. The models are trained for generating reasonable predictions as a response to now feature vector when a subject is going through identification. This project is therefore based on supervised learning.

Models are trained to predict when new inputs are given with classification algorithms. The target functions $y = f(x)$ is unknown and represents the correct predictions. A hypothesis function $h(x)$ approximates the unknown target function. The goal of the learning process is to find the hypothesis function that best approximates the unknown target function [28].

Obtaining a hypothesis that fits the future data best is desirable. To test the approximation of a hypothesis function, the function must be tested with unseen data. One way to estimate the accuracy of a classifier is by using a method called k-fold cross-validation. This method splits the dataset into training data and test data. The data is first split into k equal subsets. Then k rounds of learning rounds are performed. For each round, $\frac{1}{k}$ of the data is held and used for testing with the remaining for training. The average test score from k round gives a better estimate, then a single score of the classifier accuracy. Most used values for cross-validation are $k = 5$ and $k = 10$, enough for obtaining estimates statistically likely to be accurate.

In this work, five different classification algorithms are utilized for finding the best training model. The description of the classifications algorithms used in this work are described below.

Support Vector Machine

The support vector machine (SVM) is a popular choice as classification algorithms in supervised learning. The data is classified by finding the hyperplane that maximizes the margin between the classes. The hyperplane is defined by vectors called support vectors. The advantage with SVM is its capability to transform to higher-dimensional space for easier separation of nonlinear data using kernel trick. SVM is, therefore flexible to represent complex function [28].

Decision Tree

Decision Tree (DT) is a tree structure resembling a flow-chart where each node indicates a test on a feature, each branch representing the result of the test and

leaf nodes representing classes or class distribution[29]. The navigation to the different nodes is based on the test. The output is predicted once the leaf node is reached.

Random forest

The random forest (RF) is an ensemble learning algorithm, which means that it generates many classifiers and aggregates their results [30]. This algorithm consists of several DT, where each gives a prediction, and the class with the most votes become the models' prediction. This concept protects each of the classifiers from their own individual errors.

k-nearest neighbors

The k-nearest neighbors (k-NN) algorithm classifies the input with the most common class among its k-neighbors. The prediction is obtained by majority voting applied over the k nearest data points. The k-NN do not train during testing; the k-neighbours with minimum distance will take part in the classification. For determining the best k-value for the given data, the algorithm is run several times with different k-values [31].

Naive Bayes

The Naive-Bayes (NB) is a probabilistic classifier based on Bayes' Theorem with the assumption that there are no dependencies amongst the features. The reason for this is to simplify the computation [29].

Chapter 3

Literature Review

3.1 Feature extraction

The research field for using EEG signals for Subject Identification is growing with many different methods. As feature selection are essential for improving both accuracy and computational efficiency, this has to be prioritized for building an effective biometric system.

A study of feature extraction and classification method were executed in [32]. The authors were studying three different classification algorithms, SVM, k-NN, and NB, employed with two different feature extraction methods, EMD and Discrete Wavelet Transform (DWT). The aim of the study was Subject identification using low-density EEG signals of resting-states data. The dataset contained recordings from 27 subjects (5 sessions with 30 instances each) using one set with 14 channels and four subsets (8, 4, 2, and 1 channel). A greedy algorithm was utilized for reducing the number of channels with a minimum loss of accuracy. EMD showed to be more robust as a technique for feature extraction of EEG during resting state, especially when the number of channels was reduced. The study also showed that linear SVM gives a higher accuracy rate for when high-density EEG-recordings are used, while Gaussian naive Bayes is better for low-density EEG-recordings.

Subject Identification based on imagined speech using EMD is presented in [33]. The EMD was used to decompose the EEG signals with the Minkowski distance for deciding the most relevant IMF for each EEG channel. Four energy features for each IMF have been computed: Instantaneous and Teager energy distribution and Higuchi and Petrosian Fractal Dimension. The dataset contained 20 subjects imagining 30 repetitions of five words in Spanish in resting state. The result from four different classifiers (random forest, SVM, naive Bayes, and k-NN) was used to compare performance. The 10-folds cross-validation gave accuracy up to 0.92 using Linear SVM.

In [34], a general methodology to determine the most effective brain rhythm for human identification is presented. Features from both the time (maximum value, standard deviation, skewness, and kurtosis) and frequency domain (FFT and Power Spectral Density (PSD) mean and maximum) were extracted from different brain rhythms and used for classification on four different neural networks. The features from the time and frequency domain were used to build the networks, with a feed-forward backpropagation algorithm for building the neural network. By comparison, beta rhythm gave the best performance with a deficient mean square error while delta rhythm gave the worst performance with a relatively higher mean square error for identifying a subject. The paper concludes with beta terms as the most efficient frequency band for human identification when using EEG signals from resting-state and problem-solving conditions.

Feature extraction based on the HHT for biometric identification with EEG signals was performed in [27]. Features were computed from the instantaneous frequency and instantaneous amplitude after taking HHT of IMFs extracted using EMD. The purposed system was tested on two datasets with different protocols with only one recording channel. One dataset with 122 subjects was acquired by users viewing a series of pictures, while the other dataset with 109 subjects was based on performing motor and imagery tasks. The Linear Discriminant Analysis (LDA) -based classifier was used for classifying instantaneous amplitude-based features. In contrast, the k-NN (3-NN) was adopted for the instantaneous frequency-based features, as this gave the best results. The average accuracies for the two datasets using only a single electrode were 0.96 and 0.99, respectively. It was also shown that lower frequency bands yield better biometric performance for both datasets. The methods must be tested on other paradigms for validation. The first dataset had only one session, while the second dataset had three sessions but separated by only a few minutes. An ideal dataset to test this method on is with multiple sessions with intervals of several days to establish stability.

In [35], a method for subject identification based on visual evoked potential (VEP) signals and neural network (NN) was proposed. A backpropagation NN was trained to identify subjects using the gamma frequency band (30-35 Hz) with the spectral power ratio of VEP signals. The dataset included 20 individuals, each recorded with 61 electrodes. Utilizing a zero-phase Butterworth digital filter and Parseval's time-frequency equivalence theorem, the gamma-band spectral-power ratio was computed. The NN classification gave an average accuracy of 0.99 across 400 test VEP patterns using a 10-fold cross-validation scheme.

In [36], event-related potentials (ERPs) as a base for an identity authentication system was tested. The dataset consisted of 26 subjects giving a feedback-related response of a P300-speller. The EMD was utilized as a feature extraction method where two IMFs where extracted from each channel. The selection of IMFs was based on the Minkowski distance. The SVM was used for classification with the ac-

curacy index computed using the 10-fold cross-validation. Greedy algorithms were used for reducing or increasing the number of channels. The accuracy of using nine channels was 0.97 for classifying 24 subjects; the accuracy decreased to 0.91 when only five channels were used.

3.2 Noise Reduction

All raw EEG-signals must go through pre-possessing before feature extraction. The reason for this is that raw EEG-signals are contaminated with electrical artifacts while recording. Artifacts occur from 50Hz or 60 Hz noises from electronics nearby the subject or from the subject itself (muscle movements, blinking, or movement of the face). Increasing the signal-to-noise ratio of EEG-signals before feature extraction by removing or filtering artifacts will improve the signal quality.

The removal of powerline noise caused by AC power supply is feasible by adding a notch filter with a narrow stop-band at 50/60 Hz as done in several experiments [37].

The known brain activities are divided into frequency bands. In [34], by removing brain rhythms not related to the given task improved signal quality by adding various filters to keep the most relevant information. The different frequency bands were extracted by first adding a low pass filter for removing line frequencies (50 Hz) and a bandpass finite impulse response (FIR) filter for filtering between 0.5 - 44 Hz. After filtering the EEG-signals, they were separated in different frequency range with AcqKnowledge software.

Butterworth filter is also a method for noise reduction of raw EEG-signals. In [35], visual evoked potential (VEP) signals were filtered to obtain the gamma-band spectral range of 30-50 Hz by using a zero-phase Butterworth digital filter. Using the zero-phase cancels the effect of the phase nonlinearity of Butter worth filtering.

In [38], Independent Component Analysis (ICA) was used for removing artifacts in EEG signals. An summery of literature review is presented in table 3.1.

Source	**[32]**	**[33]**	**[34]**	**[27]**	**[35]**	**[36]**
No. Chs.	14, 8, 4, 2, 1	14 4	-	1	61	14, 9
No. Subj.	27	20	3	122/109	20	26
Paradigm	Resting	Resting	Solve problem mental with eyes closed	VEP and imaginary task	VEP	ERP
Pre-proc.	On DWT	No	yes	no	yes	no
Artifacts removing	-	-	yes	no	yes	no
Feature extraction method	EMD, DWT	EMD	Frequency bands	EEMD	Butterworth digital filter	EMD
Features	energy, fractal	energy, fractal	statistics, frequency	HHT-based	spectral power rations	energy, fractal
Classification	SVM, k-NN, NB	RF, NB, SVM, k-NN	Artificial -NN	LDA, k-NN	NN	EMD
Accuracy	0.91	0.92	1.0	0.96	0.99	0.97

Table 3.1: Summary of literature review

Chapter 4

Materials and methods

4.1 Datasets

Two datasets containing different neuro-paradigms are utilized for comparing feature extraction methods — one dataset containing subjects in resting-state and the other with subjects executing a cognitive task.

4.1.1 Resting-state

The dataset was acquired from an experiment where the participants were relaxed with eye-closed [39]. The dataset consists of EEG signals from eight subjects (five males and three females, range: 24-28). The signals were collected using the Emotiv Epoc+ headset with 14 channels with a sampling rate of 128 Hz [40]. The dataset contains three sessions for each subject, with each session containing one instance of 7000 samples. Each instance is then divided into a total of 20 sub-instances, each containing 256 samples. The specs of the headset are presented in Table 4.1.

4.1.2 Event related potential

The dataset is from a Brain Computer Interface Challenge, proposed by the IEE Neural Engineering Conference [41]. The dataset contains EEG signals from people trying to spell a word by paying attention to visual stimuli; this is known as the "P300-Speller" paradigm used in Brain Computer Interface. The P300-Speller paradigm uses EEG and P300 responses to select items displayed on a screen.

The dataset was collected from an experiment where each subject was presented letters and numbers to spell words. Items in a group were flashed on the screen in random order for selecting a letter of a word. The selected letter by an online algorithm could either be correct or wrong. The feedback response given by the

observing subject will be a P300 response of the selected item lasting for 1.3 s. See Figure 4.1 for the protocol design using a P300-speller.

This experiment was conducted on 26 subjects (13 male, range 10-37, mean age 28.8 ±5.4 (SD)). The EEG signals were recorded using 56 passive Ag/AgCl EEG sensors that followed the extended 10-20 system. Each subject went through five sessions, where each session containing 60 instances with 260 samples each.

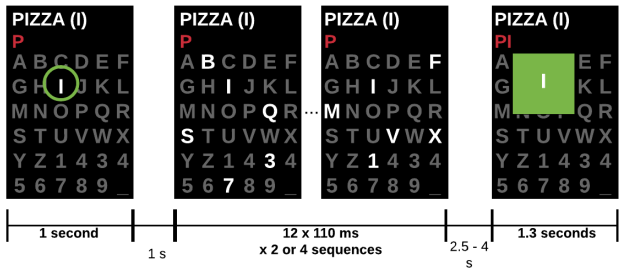

Figure 4.1: Protocol design using P300-speller [42]

Dataset	Subjects	Session	Instances	Channels	Other
Resting-state	8	3	1 Subinstances: 10, 20	14	Bandwidth 0.16 Hz - 43 Hz Digital notch filter (50 Hz / 60 Hz) Digital 5th order Sinc filter
ERP	26	5	10, 20, 40, 60	56	-

Table 4.1: Summary of the datasets used in this project.

4.2 Pre-processing

Raw EEG signals typically consist of electrical artifacts, as mentioned earlier. It is, therefore, necessary to pre-process EEG signals to remove or reduce artifacts and noise for improving the SNR.

One method for improving the SNR is by filtering the signal. Powerline noise caused by AC power supply can be suppressed by applying a notch filter (band-stop filter) with a narrow stopband at 50 Hz or 60 Hz depending on the AC frequency [5]. This can also be done using high or low pass filtering depending on what frequencies are of interest for the experiment.

Another method for removing high-frequency noise is by utilizing EMD. Extracting the first IMF from the original signal can increase the SNR as the first IMF contains the high-frequency mode of the signal. However, depending on the experiment, the first IMF can hold on useful information that should not be removed. There is no defined solution for this process. Therefore, a combination of pre-processing methods, as well as no pre-processing, are explored. Regardless,

pre-processing is a crucial step for obtaining high-quality data. An advanced review is out of the scope for this work.

4.2.1 Preprocessing of resting-state

EEG collection headsets regularly introduce constant noise in the recorded signals [39]. The headset used for obtaining the resting-state data introduced a DC offset of 4200 muV. The signals are therefore pre-processed by subtracting the constant DC offset from the raw EEG signals. No other pre-processing methods are applied to this dataset since the Empotiv Epoc+ headset used has built-in bandwidth filters.

4.2.2 Preprocessing of ERP

By taking the FFT of the raw signals from the ERP signals, the signals affected by the powerline noise is visible as all the subjects produce a high-frequency response at 50 Hz shown in Figure 4.2. A notch filter with stopband at 50 Hz is therefore applied to the signals before any further work. According to [43], the frequency band of a P300 response range from 0.5 Hz - 70 Hz. A band-pass filter from 0.5 Hz -70 Hz is therefore applied to the signals for removing frequencies over and under this range. The Infinite impulse response (IIR) Butterworth band-pass filter used has a maximally flat magnitude in the passband [5].

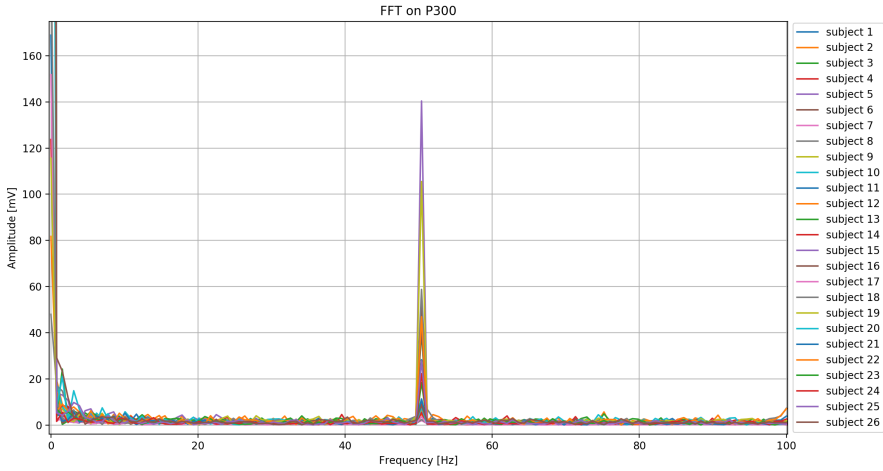

Figure 4.2: FFT plot of raw EEG signals from ERP data. The powerline noise is visible at 50 Hz with high amplitude value for each subject.

4.3 Decomposition

Two signal decomposition methods are utilized before feature extractions. One method is the EMD, and the other is the IIR Butterworth bandpass and highpass filter for extracting the brain frequency bands. How the signal decomposition methods work as a basis for feature extraction will be examined on both neuro-paradigms.

4.3.1 Decomposing with Empirical Mode Decomposition

This method is based on [42]. The EMD is used for extracting IMFs using cubic spline for interpolation on each channel in a dataset. Depending on the signal size, a different number of IMFs are obtained. The most relevant IMFs are therefore chosen using the Minkowski distance as proposed [44]. For feature extractions, it is important to use the same number of IMFs for all instances. Therefore, the first two IMFs are used from all channels, as this was the minimum number of relevant IMFs in all channels. This was the case for both datasets. See Figure 4.3 for illustration of signal decomposition using EMD.

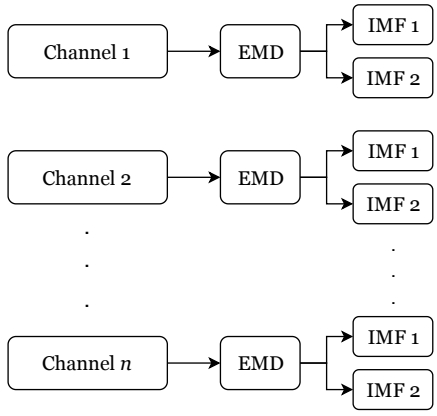

Figure 4.3: EMD applied as signal decomposition method on each EEG channel

EMD is a data-driven method, which means that IMFs are extracted if the two conditions mentioned in section 2.4.2 are fulfilled. For some specific instances, these conditions do not fulfill when using EMD on the dataset based on ERP. For these instances, IMFs are not extracted from the signals, and no feature extractions are executed. The number of feature extractions must be equal in all sessions for all subjects to be able to compare the extracted features. Therefore, instances 15, 21, 23, and 45 are removed from the dataset containing ERP as these do not fulfill the conditions for IMF when EMD is utilized. The number of instances used for training and testing are 10, 19, 37, and 56.

4.3.2 Decomposition with frequency bands

Four Butterworth bandpass filters and one highpass filter are applied on each channel to extract the brain frequency bands. Each band contains EEG signals with frequency range mentioned in table2.1. See Figure for illustration of signal decomposition with frequency bands. 4.4.

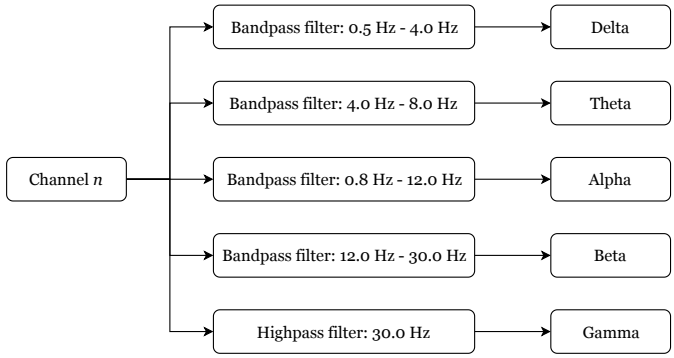

Figure 4.4: Decomposing signals into frequency bands applied on each EEG channel

4.4 Feature extraction

After signal decomposition is applied to EEG signals, features are then extracted. The following features presented in Table 4.2 are extracted from each dataset:

Features	Extracted features
Energy	instantaneous energy and Teager energy
Fractal	Petrosian and Higuchi fractal dimension
HHT-based	Marginal frequency and mean instantaneous amplitude
Statistical	min, max, mean, median, variance, standard deviation, kurtosis, skew

Table 4.2: Features extracted from ERP and resting-state data.

The method proposed in [42] shows that feature set containing energy and fractal features gives high accuracy for classification. The behavior of other feature combinations are examined by testing three different features sets on both datasets, as presented in Table 4.3.

Features sets	Features
Set 1	Energy, fractal, HHT-based
Set 2	Energy, marginal frequency (from HHT-based features)
Set 3	Statistical

Table 4.3: Features sets used on ERP and resting-state data after decomposition.

4.5 Classification using feature sets

An overview of the signal decomposition methods and classification used for obtaining accuracy from subject identification is described in four steps:

1. *Decompose the individual EEG signals*

From each dataset, signals are decomposed using EMD (two IMFs) and frequency bands (five brain frequency bands) separately. The EMD algorithm utilized in this project is from the PyEMD package [45]. The IIR Butterworth bandpass and highpass filters are created using the open-source Python Scipy Signal processing package.

The dataset containing ERP data has been used in different papers [42], [36], for examining the accuracy when the number of channels are reduced. Based on these papers, both signal decomposition methods are applied to three different sets of channels when using ERP data. This is done for examine the evolution of accuracy when reducing the number of channels and instances. The dataset is used with 56 channels (all the channels), 32 channels (used in [42]), and seven channels (used in [36]).

2. *Create a feature vector for each instance*

For each instance, the three feature sets presented in table 4.3 are computed on each channel and combined to form in a feature vector. An illustration for obtaining feature vector using EMD as a basis for feature extraction is presented in Figure 4.5.

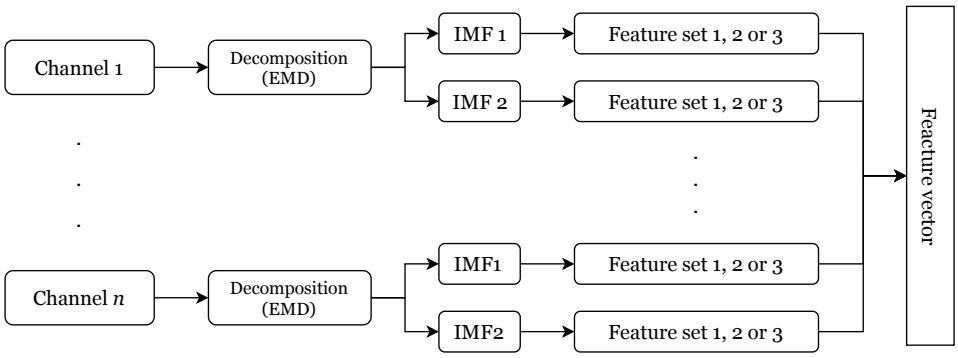

Figure 4.5: Illustration of feature extraction using EMD as basis. Feature vector are created for each instance by combining the the results obtained from each channel.

3. *Classify the features vectors and obtain accuracy for each classifier*

The feature vectors obtained from each instance are combined to one single feature vector per session. The complete feature vector is then used as input for the classifiers. Depending on what feature sets are used, the size of the feature vector will vary. An instance containing seven channels, two IMFs, and statistical features (8 features), gives a feature vector with size 112 ($7 \times 2 \times 8 = 112$).

4. *Select classifier with highest accuracy*

Machine-learning models in the supervised form are created using 10-fold cross-validation for obtaining model accuracy. The classification algorithm DT, RF, k-NN, SVM, and NB are utilized for creating models. The classification models are obtained using different parameters for each classifier for finding the best parameters. The parameters used are:

1. RF, DEPTHS = [2, 3, 4, 5, 6]
2. k-NN, NEIGHBORS = [2, 3, 4, 5, 6, 7, 8, 9, 10]
3. SVM, kernels = [linear, radial basis function, sigmoid polynomial]

For the NB classifier, the GaussionNB from the sckit-learn library with its default parameters utilized in this project. Scikit-Learn is an open-source machine learning library in Python containing several built-in classification algorithms. An illustration of the whole from pre-processing to classification is illustrated in Figure 4.6.

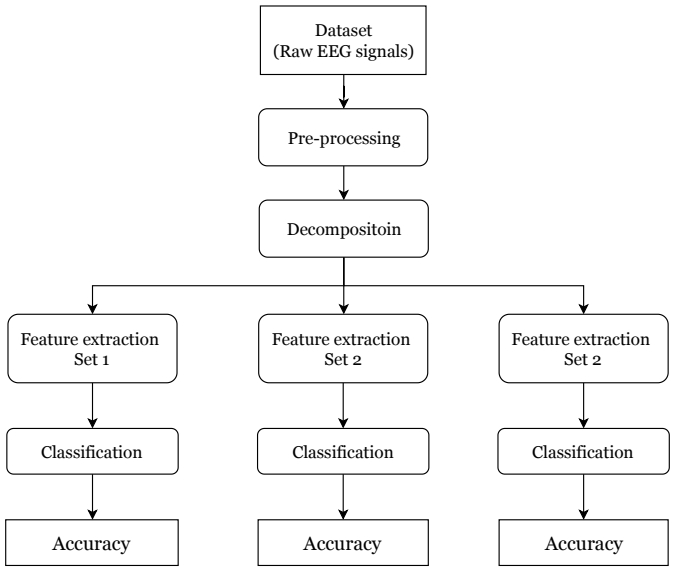

Figure 4.6: Classification process executed on each data set

Chapter 5

Results and discussion

5.1 Pre-processed data

The ERP data after pre-processing is depicted in Figure 5.1. The powerline noise at 50 Hz is suppressed, and the high-frequency value at 0 Hz is also suppressed after the bandpass filtering from 0.5 Hz - 70.0 Hz. The high-frequency value at 0 Hz before pre-processing is from the offset in the signals. The high amplitude values at 50 Hz are still visible for two of the subjects. These subjects had the highest amplitude value at 50 Hz before pre-processing and need a more narrow stopband for suppressing the 50 Hz noise.

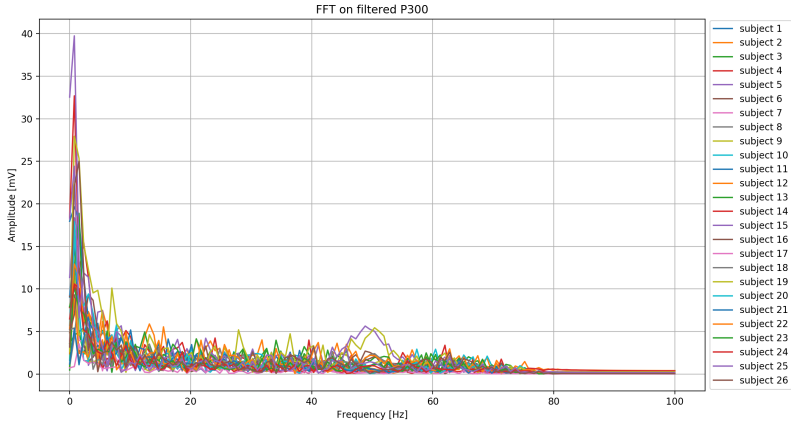

Figure 5.1: FFT of ERP data after pre-processing shows the 50 Hz noise from powerline and offset at 0 Hz are suppressed.

5.2 Classification with feature extraction

The combination of signal decomposition method and feature extraction providing highest validation value is of interest, and the following experiments are proposed:

1. Recreating methods used in [32] from the literature review.

2. Classification using all three feature sets separately on each decomposition method on both datasets.

All the experiments follow the procedure described in section 4.5. The accuracy and validation are used for investigating the performances. For the resting-state data, the models are trained using the first two sessions, and for the ERP data, the first four sessions are used for training the models. The validation for both datasets is obtained by testing the last session from each dataset on the trained model with the highest accuracy (session three for resting-state data and session five for ERP-data).

The first experiment aims to investigate which classification algorithm gives the highest accuracy value for low- and high-density EEG-recordings on different neuro-paradigms. The second experiment aims to investigate which signal decomposition method used as a basis for feature extraction gives the highest validation with different feature sets.

5.2.1 Recreating literature review

According to [32], SVM gives a higher accuracy rate when high-density EEG-recordings are used, while Gaussian naive Bayes is better for low-density EEG-recordings on resting-state. The method executed in [32] is recreated to investigate if the same results can be obtained using ERP data. The method is replicated by feature extraction with energy and fractal features. Both signal decomposition methods are utilized as basis for feautre extraction for comparison.

Table 5.1 and 5.2 presents the accuracy results obtained for ERP data and resting-state data, respectively. The highest overall accuracy is obtained with both datasets. In the case of ERP data, the highest accuracy obtained was 1.0, using EMD as the basis for feature extraction with all channels and the first ten instances. For the resting-state dataset, the highest accuracy of 1.0 was obtained using frequency bands with 20 instances.

The validation values for ERP and resting-state data are presented in table 5.3 and 5.4, respectively. The highest validation value obtained for ERP data using energy and fractal features was 0.92 using frequency bands as the basis. For the resting-state data, highest validation value obtained was 0.98 also using frequency bands as the basis for energy and fractal features.

Methods	Pre-processing	Channels	Instances			
			10	**19/20**	**37/40**	**56/60**
EMD	Yes	7	0.75	0.76	0.77	0.77
		32	0.91	0.91	0.92	0.93
		56	0.91	0.94	0.94	0.95
	No	7	0.93	0.94	0.93	0.94
		32	0.99	0.99	0.98	0.98
		56	**1.0**	0.99	0.99	0.99
Frequency bands	Yes	7	0.76	0.83	0.85	0.87
		32	0.88	0.92	0.94	0.95
		56	0.89	0.93	0.96	0.96
	No	7	0.92	0.94	0.95	0.96
		32	0.98	0.98	0.98	0.98
		56	**0.99**	0.98	0.99	0.99

Table 5.1: Accuracy of ERP data using energy and fractal features with EMD and frequency bands as basis for feature extraction. 10, 19, 37 and 56 are number of instances used when EMD is utilized.

Methods	Channels	Pre-processing	Instances	
			10	**20**
EMD	14	Yes	0.97	0.99
		No	0.97	**0.99**
Frequency bands	14	Yes	0.97	1.0
		No	0.97	**1.0**

Table 5.2: Accuracy of resting-state data using energy and fractal features with EMD and frequency bands as basis for feature extraction.

Methods	Pre-processing	Channels	Instances			
			10	**19/20**	**37/40**	**56/60**
EMD	Yes	7	0.52	0.54	0.62	0.64
		32	0.74	0.73	0.76	0.79
		56	0.76	0.80	0.82	0.82
	No	7	0.75	0.77	0.79	0.79
		32	0.89	0.92	0.90	0.94
		56	**0.89**	0.86	0.88	0.89
Frequency bands	Yes	7	0.58	0.66	0.70	0.74
		32	0.71	0.75	0.81	0.89
		56	0.75	0.79	0.86	0.88
	No	7	0.78	0.81	0.84	0.85
		32	0.83	0.90	0.89	**0.92**
		56	0.89	0.91	0.90	0.91

Table 5.3: Validation of ERP data using energy and fractal features.

Methods	Channels	Pre-processing	Instances	
			10	**20**
EMD	14	Yes	0.88	0.95
		No	0.88	**0.95**
Frequency bands	14	Yes	0.96	0.98
		No	0.96	**0.98**

Table 5.4: Validation of resting-state dataset using energy and fractal features.

The evolution of accuracy using different classification algorithms with EMD as the basis using 20 instances is presented in Figure fig. 5.2. The classification algorithm giving the highest accuracy for both low-density and high-density records on ERP data is the linear SVM. This result also yields for other numbers of instances.

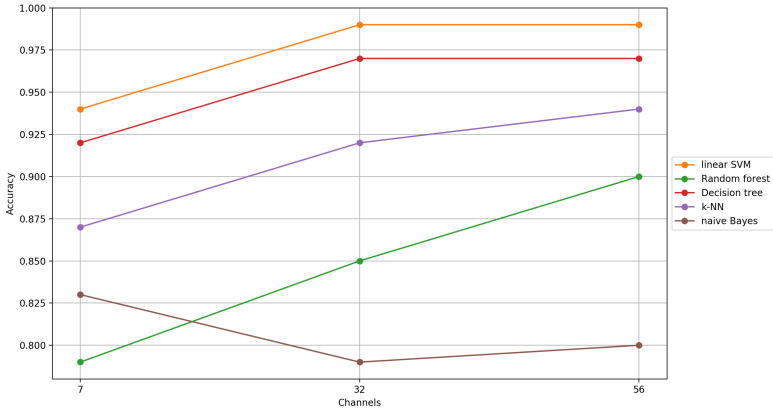

Figure 5.2: Evolution of accuracies obtained using 20 instances with EMD as basis on ERP data, with Linear SVM, random forest, decision tree, k-NN, and Gaussion naive Bayes.

5.2.2 Classification using feature sets

After all training models were obtained for each dataset, the model with the highest accuracy was validated by testing the model with unseen data. This setup of only testing the models with the last session does not resemble the real-world application. In a real-life scenario, the trained model will most likely contain only one session of training and should be taken into consideration when evaluating.

Results resting-state dataset

The results with validation values obtained using frequency bands and EMD as basis for feature extractions are presented in table 5.5 and table 5.6, respectively. The highest overall validation value 0.96 was obtained using frequency bands as the basis for feature extraction, with the feature set containing energy, fractal, and HHT-based features. The maximum number of 20 instances were used with no pre-processing. The highest validation value obtained using EMD as the basis for feature extracting was 0.86, using feature set containing energy, fractal, and marginal frequency from HHT-based features using 20 instances.

Features set	Pre-processing	10 instances	20 instances
Set 1	Yes	0.86	0.96
	No	0.86	**0.96**
Set 2	Yes	0.86	0.91
	No	0.86	0.91
Set 3	Yes	0.81	0.92
	No	0.81	0.92

Table 5.5: Validation of resting-state data using frequency bands as basis for feature extraction

Features sets	Pre-processing	10 instances	20 instances
Set 1	Yes	0.70	0.82
	No	0.70	0.82
Set 2	Yes	0.74	**0.86**
	No	0.74	0.86
Set 3	Yes	0.66	0.83
	No	0.74	0.86

Table 5.6: Validation of resting-state data using EMD as basis for feature extraction

Results of ERP dataset

The validation value obtained using frequency bands and EMD as the basis for feature extractions on ERP data are presented in 5.7 and 5.8, respectively. The validation table for the ERP data is not complete when using EMD as a basis. The overall highest validation value from the available results obtained was 0.86 using frequency bands as the basis for feature extraction with the feature set containing energy, fractal, and marginal frequency from HHT-based features with 56 instances. The highest validation value obtained using EMD as the basis was 0.78, with statistical features using 56 instances.

Features	Channels	Pre-processing	Instances			
			10	**20**	**40**	**60**
Set 1	**7**	Yes	0.43	0.45	0.44	0.48
		No	0.55	0.54	0.62	0.56
	32	Yes	0.43	0.48	0.49	0.53
		No	0.56	0.61	0.69	0.58
	56	Yes	0.42	0.60	0.51	0.58
		No	0.53	0.54	0.64	0.69
Set 2	**7**	Yes	0.40	0.41	0.47	0.48
		No	0.52	0.53	0.65	0.58
	32	Yes	0.62	0.49	0.53	0.54
		No	0.60	0.67	0.71	0.66
	56	Yes	0.44	0.50	0.47	0.55
		No	0.55	0.64	0.64	0.70
Set 3	**7**	Yes	0.50	0.63	0.70	0.74
		No	0.64	0.68	0.75	0.74
	32	Yes	0.58	0.70	0.78	0.81
		No	0.77	0.82	**0.86**	0.85
	56	Yes	0.59	0.70	0.80	0.82
		No	0.75	0.67	0.80	0.83

Table 5.7: Validation of ERP dataset using frequency bands as basis for feature extractions.

Features	Channels	Pre-processing	Instances			
			10	**19**	**37**	**56**
Set 1	**7**	Yes	0.4	0.37	0.38	0.45
		No	0.5	0.55	0.60	0.57
	32	Yes	0.46	0.46		
		No		0.67	0.57	
	56	Yes	0.47	0.44	0.49	
		No		0.59	0.56	0.62
Set 2	**7**	Yes	0.36	0.37	0.44	
		No	0.48	0.56	0.58	
	32	Yes	0.41	0.51		
		No	0.60	0.57	0.64	0.64
	56	Yes	0.48	0.47	0.50	
		No		0.52	0.56	0.59
Set 3	**7**	Yes	0.42	0.48	0.49	0.52
		No	0.59	0.59	0.69	0.71
	32	Yes	0.55	0.61	0.68	0.69
		No	0.70	0.77		
	56	Yes	0.58	0.65	0.70	0.75
		No	0.70	0.62	0.77	0.78

Table 5.8: Validation of ERP dataset using EMD as basis for feature extractions.

Chapter 6

Discussion

6.1 Recreating literature review

The first experiment with energy and fractal features for feature extraction shows promising results for both ERP and resting-state data. This combination of features gave high accuracy for both resting-state data and ERP data. For both cases, using only the first ten instances was enough for obtaining high accuracies. The number of instances used for classification can influence accuracy. A higher number of instances can contain redundant features and reduce accuracy. By observing the evolution of validation value with the increasing number of instances, is it shown that higher validation values can be obtained with a lower number of instances. Further analysis is needed to determine the importance of instances. The resting-state data gave the highest overall accuracy. The reason for this could be the few numbers of subjects in the dataset. In [32], the same paradigm was used with more subjects (27 subjects in comparison to 8).

From recreating the method used in [32], it is shown that the best classification algorithms used for low-density and high-density EEG-recording on resting-state data do not match for ERP data. The linear-SVM algorithm gives highest accuracy for both low-density and high-density data when ERP data is analyzed. It is important to mention that the majority of the high accuracy values using different classifiers were obtained using the frequency bands as a signal decomposition method.

6.2 Classification using feature sets

For the second experiment with both datasets, the highest overall validation value was obtained using the frequency bands as the basis for feature extraction. The validation table using EMD as the basis for feature extraction was incomplete, primarily where energy, fractal, and HHT-based features were extracted. The reason for this was extremely long response time from the SVM. The 10-cross-validation was employed to calculate the model accuracy, which increases the calculation time compared to 5-cross validation since the trained model will be tested on more sets of test groups. Another reason for the long computation time could be the use of selected features. Adding HHT-based features to energy and fractal features could make it more challenging for classification. A combination of feature selection and testing with different classification algorithms should be considered in further work.

6.3 Overall discussion

For all the experiments, higher accuracies were obtained on both signal decomposition methods when no pre-processing was applied to the signals. Pre-processing used in this work may have removed the unique characteristics from the different EEG signals, making it difficult to classify the data. For instance, high-amplitude peaks at 50 Hz caused by the powerline were different for most of the subjects. This could be an important feature for differentiating EEG signals from different subjects. Using other pre-processing methods such as ICA mentioned in the literature review could improve the SNR without removing important information from the signals and give higher accuracy and validation.

When examining the accuracy with different sets of channels used on ERP data, the highest value is obtained when all channels are used. Lowering the number of channels is of interest when subject identification using EEG signals are applied in real-time application. The highest validation value obtained using seven channels was 0.85, with frequency bands as the basis for feature extraction with the feature set containing energy and fractal features. The highest validation value obtained using 32 channels was 0.92 using frequency bands as a feature extraction base with statistical features; this was also the overall highest validation value using EMD as the basis on ERP data. This shows the possibility of obtaining high accuracy using reduced data. In this work, the channel selection was based on different papers using the same dataset. For future work, methods for channel selection before classification should be applied for finding the best channel for given neuro-paradigms. The use of greedy algorithms for channel selection purposed in [32] is on method for this.

The difference in the number of subjects in the datasets makes it difficult to compare which neuro-paradigm suits best for subject identification. Validations values are higher with resting-state, but increasing the number of subjects could

make the validation value decrease. Further work should contain datasets with a higher number of subjects for providing more secure results.

Feature selection was not applied in this work. By adding feature selection before classification, redundant features can be removed, which could increase the accuracy and validation. Due to limited time, feature selection was not prioritized and will be part of future work.

Using frequency bands appears to be more secure then EMD in this project. In the case of EMD as a signal decomposition method, not all instances were utilized as they did not meet the conditions for extracting IMFs. The reduced number of instances could be a reason for the lower validation values using EMD. Therefore, frequency bands seem to be a better option as a feature extraction basis because it is not data-driven, nor does it need any predefined parameters. This method can also be used by only extracting relevant frequency bands. Resting-state usually works the lower frequency area; the brain frequency bands delta, alpha, and maybe beta may be of interest. By including feature selection to this method, this could be a suitable method for extracting relevant features from EEG-signals.

Chapter 7

Conclusion

In this project, a comparison of EMD and frequency bands for subject identification using EEG-signals from resting-state and ERP data has been presented. The use of a reduced number of recording channels and fewer instances were also examined.

Pre-processing by using a bandpass filter and notch filter removed necessary information from recorded EEG-signals, which leed to lower accuracy and validation values. The signal decomposition methods alone gave high accuracy and validation values for the classification.

The purposed method from [32] were recreated to examine if linear SVM and Gaussian naive Bayes gives higher accuracies for lower- and higher-density EEG-recordings on other neuro-paradigms than resting-state. The results from utilizing the purposed method on ERP data showed that linear SVM gives highest accuracy value for both lower- and higher-density EEG-recording.

EMD and frequency bands were used to decompose the EEG-signals. Various types of features such as energy, fractal, and HHT-based were then calculated from the decomposed signals to obtain feature vectors. The vectors were then used as input for several classification algorithms to train models. The model with highest accuracy was used for validation with unseen data. The validations of models using EMD as a basis for feature extraction were incomplete, as some signals from the ERP data did not meet the conditions for extracting IMFs.

The overall highest validation accuracy obtained in this project was 0.98. This was obtained using frequency bands as a basis for feature extraction with energy and fractal features on resting-state data.

7.1 Future work

A more broad study of the currently used signal decomposition methods is required for high classification accuracy and validation for using a reduced number of features. Future work will, therefore, include feature selection for removing redundant features. This will reduce the computation time in classification and represent the EEG signals with more relevant features for more accurate identification.

Channels selection and feature selection will also be an essential part of feature work as is also can reduce the computation and make it more appropriate for real-time application.

A practical implementation of a subject identification system based on EEG signals will be the final part of future work.

Bibliography

[1] A. K. Jain, A. Ross, and S. Prabhakar. An introduction to biometric recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(1):4–20, Jan 2004.

[2] A. K. Jain, A. Ross, and U. Uludag. Biometric template security: Challenges and solutions. In *2005 13th European Signal Processing Conference*, pages 1–4, Sep. 2005.

[3] A. K. Jain, A. Ross, and K. Nandakumar. An introduction to biometrics. In *2008 19th International Conference on Pattern Recognition*, pages 1–1, Dec 2008.

[4] Anil K. Jain, Ruud Bolle, and Sharath Pankanti. *Biometrics: Personal Identification in Networked Society*. Kluwer Academic Publishers, Norwell, Massachusetts, 1999.

[5] Qiong Gui, Maria V. Ruiz-Blondet, Sarah Laszlo, and Zhanpeng Jin. A survey on brain biometrics. *ACM Comput. Surv.*, 51(6):112:1–112:38, February 2019.

[6] Q. Gui, Z. Jin, and W. Xu. Exploring eeg-based biometrics for user identification and authentication. In *2014 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pages 1–6, Dec 2014.

[7] James L Wayman. Fundamentals of biometric authentication technologies. *International Journal of Image and Graphics*, 1(01):93–113, 2001.

[8] Jonathan Wolpaw and E.W. Wolpaw. *Brain-Computer Interfaces: Principles and Practice*. 01 2012.

[9] Herbert H Jasper. The ten-twenty electrode system of the international federation. *Electroencephalogr. Clin. Neurophysiol.*, 10:370–375, 1958.

[10] Luis Luis and Jaime Gomez-Gil. Brain computer interfaces, a review. *Sensors (Basel, Switzerland)*, 12:1211–79, 12 2012.

[11] Subha Dharmapalan Puthankattil, Paul Joseph, U Rajendra Acharya, and Choo Lim. Eeg signal analysis: a survey. *Journal of medical systems*, 34:195–212, 04 2010.

[12] George H Klem, Hans Otto Lüders, HH Jasper, C Elger, et al. The ten-twenty electrode system of the international federation. *Electroencephalogr Clin Neurophysiol*, 52(3):3–6, 1999.

[13] J Craig Henry. Electroencephalography: basic principles, clinical applications, and related fields. *Neurology*, 67(11):2092–2092, 2006.

[14] Ali Haider and Reza Fazel-Rezai. Application of p300 event-related potential in brain-computer interface. In Phakkharawat Sittiprapaporn, editor, *Event-Related Potentials and Evoked Potentials*, chapter 2. IntechOpen, Rijeka, 2017.

[15] Dimitris G Manolakis and Vinay K Ingle. *Applied digital signal processing: theory and practice*. Cambridge University Press, 2011.

[16] Norden E Huang, Zheng Shen, Steven R Long, Manli C Wu, Hsing H Shih, Quanan Zheng, Nai-Chyuan Yen, Chi Chao Tung, and Henry H Liu. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 454(1971):903–995, 1998.

[17] Norden Eh Huang. *Hilbert-Huang transform and its applications*, volume 16. World Scientific, 2014.

[18] Y. Gao, G. Ge, Z. Sheng, and E. Sang. Analysis and solution to the mode mixing phenomenon in emd. In *2008 Congress on Image and Signal Processing*, volume 5, pages 223–227, May 2008.

[19] Zhaohua Wu and Norden E Huang. Ensemble empirical mode decomposition: a noise-assisted data analysis method. *Advances in adaptive data analysis*, 1(01):1–41, 2009.

[20] J. Jebaraj and R. Arumugam. Ensemble empirical mode decomposition-based optimised power line interference removal algorithm for electrocardiogram signal. *IET Signal Processing*, 10(6):583–591, 2016.

[21] Girish Chandrashekar and Ferat Sahin. A survey on feature selection methods. *Computers & Electrical Engineering*, 40(1):16–28, 2014.

[22] J. M. O. Toole, A. Temko, and N. Stevenson. Assessing instantaneous energy in the eeg: A non-negative, frequency-weighted energy operator. In *2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 3288–3291, Aug 2014.

[23] F. Finotello, F. Scarpa, and M. Zanon. Eeg signal features extraction based on fractal dimension. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4154–4157, Aug 2015.

[24] Cindy Goh, Brahim Hamadicharef, Goeff Henderson, and Emmanuel Ifeachor. Comparison of fractal dimension algorithms for the computation of eeg biomarkers for dementia. 2005.

[25] Tomoyuki Higuchi. Approach to an irregular time series on the basis of the fractal theory. *Physica D: Nonlinear Phenomena*, 31(2):277–283, 1988.

[26] C. F. Vega and J. Noel. Parameters analyzed of higuchi's fractal dimension for eeg brain signals. In *2015 Signal Processing Symposium (SPSympo)*, pages 1–5, June 2015.

[27] S. Yang and F. Deravi. Novel hht-based features for biometric identification using eeg signals. In *2014 22nd International Conference on Pattern Recognition*, pages 1922–1927, Aug 2014.

[28] Stuart J Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,, 2016.

[29] Ahmad Ashari, Iman Paryudi, and A Min. Performance comparison between naïve bayes, decision tree and k-nearest neighbor in searching alternative design in an energy simulation tool. *International Journal of Advanced Computer Science and Applications*, 4, 12 2013.

[30] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.

[31] Onel Harrison. Machine learning basics with the k-nearest neighbors algorithm, 2018. Last accessed 10 December 2019.

[32] Luis Alfredo Moctezuma and Marta Molinas. Subject identification from low-density eeg-recordings of resting-states: A study of feature extraction and classification. In *Future of Information and Communication Conference*, pages 830–846. Springer, 2019.

[33] Luis Moctezuma and Marta Molinas. *EEG-Based Subjects Identification Based on Biometrics of Imagined Speech Using EMD*, pages 458–467. 12 2018.

[34] M. M. Hasan, M. H. A. Sohag, M. E. Ali, and M. Ahmad. Estimation of the most effective rhythm for human identification using eeg signal. In *2016 9th International Conference on Electrical and Computer Engineering (ICECE)*, pages 90–93, Dec 2016.

[35] R. Palaniappan. Method of identifying individuals using vep signals and neural network. *IEE Proceedings - Science, Measurement and Technology*, 151(1):16–20, Jan 2004.

[36] Luis Moctezuma and Marta Molinas. Event-related potential from eeg for a two-step identity authentication system. 07 2019.

[37] Qiong Gui, Maria Blondet, Sarah Laszlo, and Zhanpeng Jin. A survey on brain biometrics. *ACM Computing Surveys*, 51:1–38, 02 2019.

[38] P. N. Kumar and H. Kareemullah. Eeg signal with feature extraction using svm and ica classifiers. In *International Conference on Information Communication and Embedded Systems (ICICES2014)*, pages 1–7, Feb 2014.

[39] SALIL S. KANHERE YUNHAO LIU TAO GU KAIXUAN CHEN XIANG ZHANG, LINA YAO. Mindid: Person identification from brain waves through atention-based recurrent neural network, 2017. Last accessed 05 December 2019.

[40] Emotiv epoc+ 14 channel mobile eeg, 2019. Last accessed 05 December 2019.

[41] Perrin Margaux, Maby Emmanuel, Daligault Sébastien, Bertrand Olivier, and Mattout Jérémie. Objective and subjective evaluation of online error correction during p300-based spelling. *Advances in Human-Computer Interaction*, 2012:4, 2012.

[42] Luis Moctezuma and Marta Molinas. Event-related potential from eeg for a two-step identity authentication system. 07 2019.

[43] Erol Başar and Aysel Düzgün. The clair model: Extension of brodmann areas based on brain oscillations and connectivity. *International Journal of Psychophysiology*, 103:185–198, 2016.

[44] Daoud Boutana, Messaoud Benidir, and Braham Barkat. On the selection of intrinsic mode function in emd method: application on heart sound signal. In *2010 3rd International Symposium on Applied Sciences in Biomedical and Communication Technologies (ISABEL 2010)*, pages 1–5. IEEE, 2010.

[45] Python implementation of empirical mode decomposition algorithm, 2017-. Last accessed 16 December 2019.