UKRAINIAN CATHOLIC UNIVERSITY

MASTER THESIS

Brain age prediction based on EEG records

Author:
Mykola KLYMENKO

Supervisor:
Dr. Vasily VAKORIN

A thesis submitted in fulfillment of the requirements
for the degree of Master of Science

in the

Department of Computer Sciences
Faculty of Applied Sciences

Lviv 2022

Declaration of Authorship

I, Mykola KLYMENKO, declare that this thesis titled, "Brain age prediction based on EEG records" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.

- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.

- Where I have consulted the published work of others, this is always clearly attributed.

- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.

- I have acknowledged all main sources of help.

- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

_____

Date:

_____

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Master of Science

Brain age prediction based on EEG records

by Mykola KLYMENKO

Abstract

Ambulatory EEG is a widespread test used in hospitals for the neurological evaluation of patients. EEG waveforms are typically reviewed by a trained neurologist to classify EEG into clinical categories. Methodologically, there is a need to classify EEG recordings automatically. Ideally, the classification models should be interpretable, able to deal with EEG of varying durations, and robust to various artifacts. We aimed to test and validate a framework for EEG classification, which satisfies such requirements by symbolizing EEG signals and adapting a method previously proposed in natural language processing (NLP). We considered an extensive sample of routine clinical EEG (n=5'850), with a wide range of ages between 0 and 100 years old. We symbolized the multi-variate EEG times series and applied a byte-pair encoding (BPE) algorithm to extract a dictionary of the most frequent patterns (tokens) reflecting the variability of EEG waveforms. To demonstrate the performance of such an approach, we used newly-reconstructed EEG features to predict the biological age of patients with Random Forest. We also correlated the relative frequencies of tokens with age. We found that the age prediction model achieved the mean absolute error of 15.9 in years. The correlation between actual and predicted age was 0.56. The most significant correlations between the frequencies of tokens and age were observed at frontal and occipital EEG channels. Our findings demonstrate the feasibility of an approach based on applying NLP methods to time series classification. Notably, the proposed algorithms could be instrumental in classifying clinical EEG with minimal preprocessing and sensitivity to the appearance of short events, such as epileptic spikes.

Acknowledgements

Contents

List of Figures

List of Tables

List of Abbreviations

BAI   Brain Age Index
BPE   Byte-Pair Encoding
CNN   Convolutional Neural Network
CPU   Central Processing Unit
EDF   European Data Format
EEG   Electroencephalography
fMRI   Functional Magnetic Resonance Imaging
GPU   Graphics Processing Unit
IQR   Inter Quartile Range
MAE   Mean Absolute Error
MD   Max Depth
MF   Minimal Frequency
ML   Machine Learning
MRI   Magnetic Resonance Imaging
NB   Number of Bins
NE   Number of Estimators
NLP   Natural Language Processing
PAA   Piecewise Aggregate Approximation
PCA   Principal Component Analysis
RF   Random Forest
SOTA   State Of The Art
WS   Window Size

Chapter 1

Introduction

1.1 Background

EEG is a record of the electrical activity of a human brain, measured as the time-varying electrical potential difference between pairs of electrodes on the scalp. In clinical practice, EEG is used as an objective test of brain function. EEG is widely available, relatively cheap, and captured in a standardized fashion making it widely available in hospitals worldwide. A typical EEG is acquired either for a short (20-30 min) period (routine EEG) or more extended periods (many hours to days) in in-patients or ambulatory patients at home.

We can highlight two major trends causing increased interest in the neuroimaging and, specifically, electroencephalography:

- Increasing population and life expectancy lead to rising of neurodegenerative diseases that were not so widespread in the last century, such as cognitive decline and dementia. These disorders appear more often with aging. That's why the healthcare system needs to understand how brain aging process affects neurodegenerative diseases, and how to effectively monitor and prevent them at early stages, since mental health is one of the main factor of maintaining a high quality of life during aging (Puvill et al., 2016)

- Aging does not affect people uniformly (Cole et al., 2018). Different aspects of the environment and individual genetics cause different aging speeds. Thus we need individual markers, which might help to identify subject-specific health characteristics, as well as a potential risk of neurodegenerative diseases.

Individual "biological age" can be viewed as a result of factors combination, such as genes, environment, lifestyle, health, and lifetime. Shifting focus towards "biological age" from "chronological age" allows developing more personalized treatment. There is a set of measurements used to assess personal biological age, such as DNA methylation status, accumulation of genetic damage, telomere length, and allostatic load, etc. (Franke and Gaser, 2019). However, since we focus more on cognitive health, there should be a marker closer to the brain function, that will allow quantifying the brain degenerative process.

Here is where "brain age" comes from. The concept of Brain age (BA) is based on the idea that a 50-year-old can "have the heart of a 20-year-old" or that the lungs of a 30-year old smoker "work like they're 80". There are various neuroimaging technics used to estimate individual BA, that we will cover later. The main idea is the following: comparing the brain age (that shows actual neurophysiological health) to the chronological age can show, how bad or good things are. This marker

on its own is a strong predictor not only for neurodegenerative dis-eases, but also physical disabilities, typical for aging people, and even mortality (Cole et al., 2018).

1.2   Motivation

Estimation of brain functions with neuroimaging tools is already broadly used and has a great potential for early detection and prevention of neurogenerative diseases. While most of the brain aging research has been done based on MRI scans datasets (Franke and Gaser, 2019), EEG has also recently appeared in the brain aging estimation field. For decades before, several studies have demonstrated that features like EEG rhythmic activity (e.g., delta, theta, alpha, beta, and gamma) change as a function of age (Al Zoubi et al., 2018). The cost of such systems might be the most important factor nowadays: EEG toolkit price starts at around $500, and MRI machines prices are ranging from $200k to $1M. Doing fieldwork with MRI / fMRI also is not going to happen, as there is no way to make such a machine truly portable. Therefore, for quicker, affordable, and accessible insights into brain function, with a tight temporal resolution, EEG is the method of choice.

Methodologically, there is a need to classify EEG recordings automatically. Ideally, the classification models should deal with EEG of varying durations, and robust to various artifacts. The algorithm also should be interpretable. Here we define interpretability as ability to perform revertible transformation, so we can trace extracted features back to original signal. In this way, the algorithm would be able to support analysis process performed by human physicians, who explore EEG recording visually and looking for local abnormalities.

One of the standing challenges in EEG classification is how to formulate input for machine learning models. At the same time, Natural Language Processing (NLP) field has various techniques for feature extraction from unstructured data, i.e. text. Usually it boils down to breaking text into repeating parts, like characters or words, part of words, or group of words, and counting their appearances in the dataset. One of the techniques is Byte Pair Encoding (BPE) algorithm: it finds parts of words – so called tokens – frequently occurred in the text (Gage, 1994).

The goal of this work is to employ BPE algorithm for clinical EEG classification. We aim to test the feasibility of this approach on age prediction task. Later the method can be transferred to other EEG classification tasks, such as finding epileptic spikes or other abnormalities.

1.3   Thesis structure

In the section 'Related Work' we cover the basics of existing neuroimaging techniques and focus especially on EEG. We review existing approaches to work with EEG data and its applications. We identify the current SOTA and research gap.

In the 'Approach to solution' we describe our pipeline of EEG transformation, feature extraction and age prediction. Here we also define evaluation metrics.

In 'Result' and 'Discussion' we present our main findings, describe the progress in the course of experiments, the final model score, and compare it to existing benchmarks.

Chapter 2

Related Work

Brain estimated age is just a part of the question. It mainly serves to calculate the difference between predicted age (PA) and chronological age (CA):

$$Difference = PA - CA$$

The terminology of this difference varies from one source to another. You can also encounter such terms as "brain predicted age difference" or brain-PAD (Cole et al., 2018), "brain age delta" (Smith et al., 2019), "brain age index" (BAI) (Hogan et al., 2021), such acronym as BrainAGE (brain age gap estimate) (Al Zoubi et al., 2018), or just "brain age gap" (Butler et al., 2020), etc. Let us refer to this term as 'delta' later in the text. Despite the diversity in naming, all these works define the delta as a difference between estimated brain age and actual, chronological age (simple subtraction). The positive delta points out to accelerated brain aging, negative one means resilience.

While mostly agreeing on this point, these sources focus more on various stages of brain age and delta analysis. Based on reviewed sources we can highlight the following questions/topics:

- What tools are used to collect raw data describing the physiology and processes of the brain?

- What measurement approaches applied to capture brain activity?

- How to decode raw neuroimaging data to extract meaningful features?

- What machine learning models are used to predict the age based on neuroimaging data

- How to estimate performace and what is the current State-of-the-Art for such predictions?

2.1 Neuroimaging techniques

EEG (electroencephalography) measures the electrical activity of our brain via electrodes that are placed on the scalp. It tells us, from the surface measurements, how active the brain is. This can be useful for quickly determining how brain activity can change in response to stimuli, and can also be useful for measuring abnormal activity, such as with epilepsy. (Noachtar and Rémi, 2009)
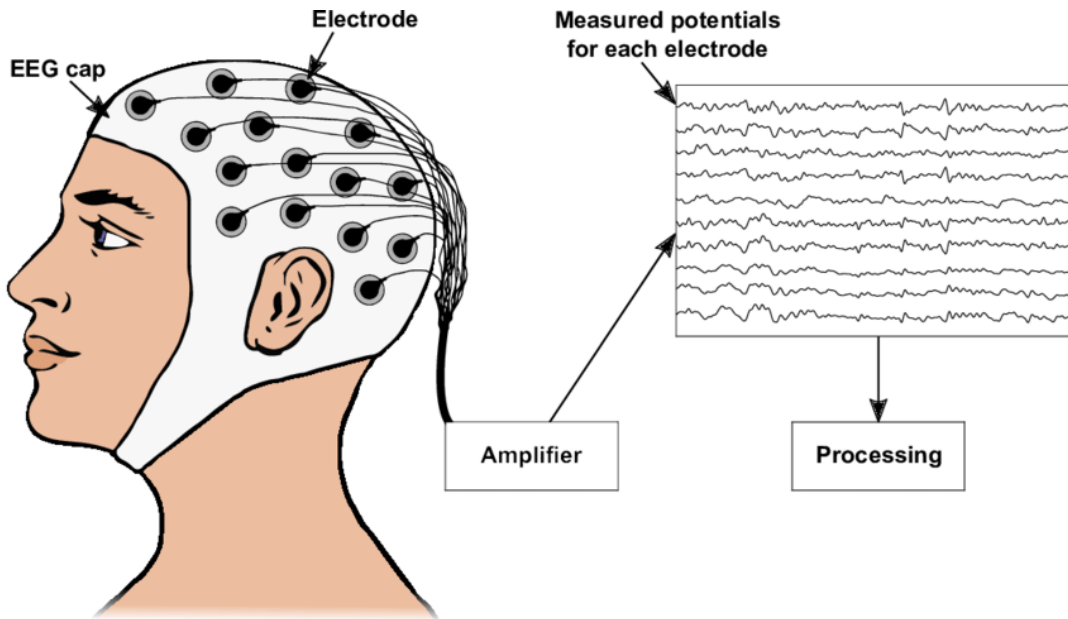
FIGURE 2.1: Sketch of recording an EEG, capturing electrical activity
on the scalp using electrodes fixated on an EEG-cap (Nagel, 2019)

The brain is an electrical system – and this electricity is generated through a network
of neurons, that send signals to each other with the help of electrical currents. The
more electrical signals, the more neuronal communication, which corresponds to
more brain activity. The electrodes of an EEG headset can't detect changes in single
neurons, but instead detect the electrical changes of thousands of neurons signalling
at the same time. A computer then receives this signal 2.1, and can generate various
maps of brain activity, with a rapid temporal resolution. (Farnsworth, 2019)

MRI (magnetic resonance imaging) provides a map of the brain – how it looks at
a set moment in time 2.2. This structural information can be useful for determining
how the sizes of certain brain areas compare across people, or if there is something
abnormal about a particular brain (a tumor for example). (Hennig et al., 2003)

As the name suggests, magnets are central to magnetic resonance imaging. The
magnetic field from the MRI interacts with the protons in our hydrogen atoms (Mills
et al., 2017). Happily, we are 70% water.

Usually, these protons are facing in random directions, but the magnetic field makes
most of them align in the same direction. For the next step, a radio pulse is emit-
ted. This also interacts with the protons, essentially turning them to the side. But,
as the radio frequency only happens for a moment, the protons relax back to their
aligned state before. As the protons relax, energy is released which can be detected
by sensors in the MRI machine. (Farnsworth, 2019)

Through some calculations (Jung and Weigel, 2013) the computer can determine
what the tissue looked like, depending on this energy that is released, and show us
an image of the tissue. MRI only shows us a static image of the brain – an anatomical
image, not of the brain's actual activity.

For **fMRI (functional magnetic resonance imaging)** the same things happen as with
MRI – the energy emitted from the relaxation of protons is measured – but the cal-
culations are instead aimed at determining how the amount of oxygenated blood
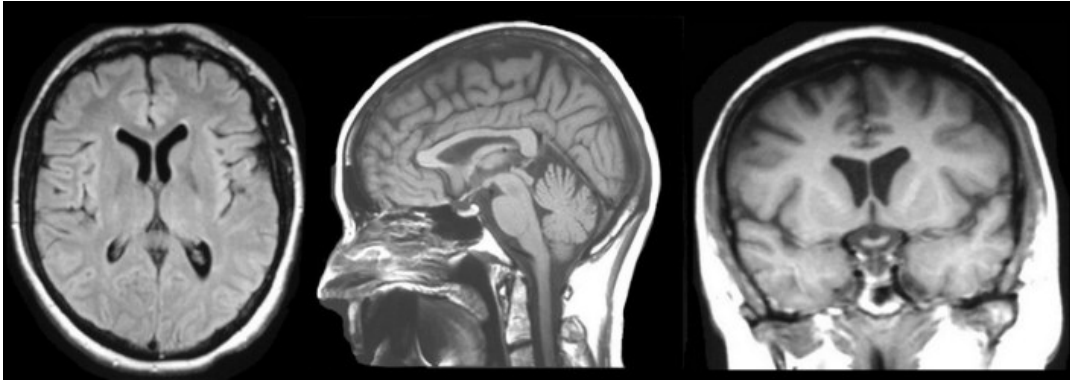
FIGURE 2.2: MRI images of a brain in all three planes: axial, sagittal and coronal (Cole et al., 2018)
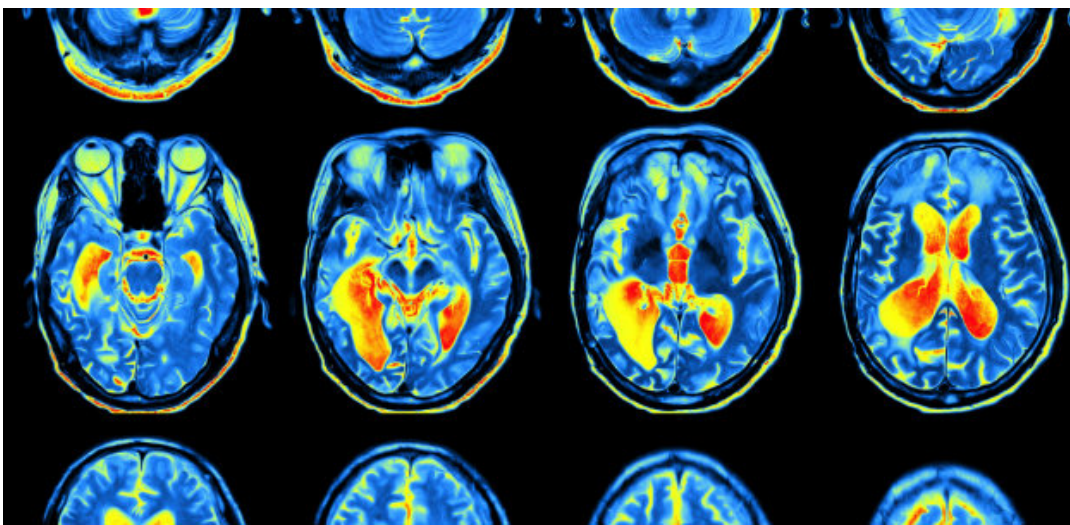
FIGURE 2.3: fMRI output sample, image credit: American Health imaging

flow changes 2.3 (Farnsworth, 2019). If there is more oxygenated blood in one part of the brain compared to others, then chances are that this brain area is more active (Hillman, 2014). This is known as the Blood-Oxygenation Level Dependent response (BOLD).

In comparison with simple MRI, fMRI gives us an image of brain activity. However, a weak point of fMRI is the temporal resolution. As it takes several seconds for the blood flow to change, and the actual recording is limited by computational factors, the data collection is slowed down. (Hennig et al., 2003)

2.2 Measurement approach

EEG recording is sensitive to the environment, since brain electroactivity quickly reacts to external stimuli, such as sound or light. Some authors rely on EEG data recorded in a short period of time, lasting 8 min. The participants were instructed to relax and keep their eyes open and fixate on a cross.(Al Zoubi et al., 2018). Also, there are attempts to collect data both with open and closed eyes and compare them to each other. Thus open-eyed EEG data allowed researchers to estimate brain age that better correlate with actual age. (Dimitriadis and Salis, 2017)

Other researchers rely on EEG recording taken during night sleep. Human sleep undergoes predictable changes with age, reflected in both overall sleep architecture and EEG oscillations/waveforms. At the level of EEG microstructure, older participants exhibit reduced slow waves during deep sleep; decreased sleep spindle amplitude, density, and duration; and less phase coupling between slow oscillations and sleep spindles (Sun et al., 2019).

Also, there is a discussion about how reliable are single time measurements for an accurate brain age estimation. The research group (Hogan et al., 2021) showed that subsequent measurements of night-sleep EEG reduce variance and error of brain age estimation. The estimated within-patient night-to-night standard deviation in BAI ('brain age index' = delta) was 7.5 years. Estimates of BAI derived by averaging over 2, 3, and 4 nights, had lower estimated standard deviations of 4.7, 3.7, and 3.0 years, respectively.

2.3   Preprocessing methods

EEG data is inherently noisy because EEG electrodes also pick up unwanted electrical physiological signals, such as the electromyogram (EMG) from eye blinks and muscles on the neck. There are also concerns about the motion artifacts occurring from cable movement and electrode displacement when the subject moves. According to the review (Craik, He, and Contreras-Vidal, 2019) of works related to solve EEG classification tasks, researchers have applied following approaches to preprocess data:

- 41% of studies did not address any specific artifact removal process,

- 29% - manual artifacts removal,

- 8% - automatic removal,

- 22% - no cleaning/removal, intentionally.

More than a quarter of the studies (26 of 90 studies) removed artifacts manually. It is indeed easy to visually identify abrupt outliers, for example when signals are lost or when intense EMG artifacts are present. However, it is difficult to identify persistent noisy channels; manual data processing is highly subjective, rendering it difficult for other researchers to reproduce the procedures. Together with the 22% of studies that did not take any actions to remove artifacts, 63% of the studies reviewed did not systematically remove EEG artifacts. The most frequent artifact removal algorithms used in the remaining 8% of the studies reviewed were independent component analysis (ICA) and discrete wavelet transformation (DWT)

2.4   Feature extraction

Feature extraction is a quintessential phase in any EEG analysis that depends on finding common features representation among EEG samples. (Teplan et al., 2002). Therefore, one of the standing challenges in EEG data analysis is how to formulate inputs. According to the review (Craik, He, and Contreras-Vidal, 2019), studies fell into three types of input formulation categories: calculated features (41%), images (20%), and the signal values (39%). The selection of input formulation relied heavily on the task and model architecture.

Calculated features. EEG data is commonly analyzed in frequency domain because they are often found to be associated with behavioral patterns. Power spectral density (PSD), wavelet decomposition, and statistical measures of the signal (i.e. mean, standard deviation) are the three most common input formulations used in the reviewed studies. Therefore, the approach is to handcraft a set of meaningful features from EEG data 2.1 (Toole and Boylan, 2017).

TABLE 2.1: Set of features extracted from EEG data

Amplitude	Total power, mean, standard deviation, skewness, kurtosis, envelope mean, and standard deviation
Peak-to-peak	Mean, median, 5th and 95th percentiles, standard deviation, the coefficient of variation and the measure of symmetry
Spectral power	Spectral power and relative power, spectral entropy (using Wiener and Shannon methods), spectral edge frequency (the cut-off frequency at which encompasses 95% of spectral power) and spectral differences between consecutive short-time spectral estimations
Connectivity	Brain symmetry index, correlation, mean and maximum of frequency at which the maximum coherence is achieved
Fractal dimension	Fractal dimension

Images input. Neural networks, especially CNN's, use spectrograms generated from the EEG data as inputs. Spectrograms are traditionally used as a postprocessing tool to visualize the data. However, CNN's ability to learn images has enabled spectrograms to be used as an input to the model. Other image formulations include creating Fourier feature maps and designing 2D or 3D grids.

Signal values. Traditionally, this approach is usually associated with particular hand-engineered time domain features, such as the power spectral density features. Neural networks promise to automatically learn complicated features from large amounts of data, prompting the idea of end-to-end learning. Feeding raw signal values directly into the neural network without hand-designed features may contribute to the practice of directly analyzing raw EEG data with deep learning. Also, it might be some statistical methods and PCA (Dimitriadis and Salis, 2017), others employ convolutional neural nets (CNN) to reduce the dimensionality and turn signal into features (Schirrmeister et al., 2017). Anyway, with this approach, we lose the interpretability of the data and results, what exactly has caused higher or lower brain age. In some cases, we can at least reconstruct it on the channel level, and see which area of the brain and scalp provides more information and explained variance.

2.5 ML algorithms for age prediction

Choosing the appropriate algorithm heavily depends on the feature extraction approach, that we have discussed in the previous subsection. In case when researchers use handcrafted features, they are free to use any classical ML algorithm that works well with tabular data: Elastic Net, Support Vector Regression, Random Forest, extreme gradient boosting tree, and Gaussian Process with Polynomial Kernel. They trained on a set of healthy individuals, so the model will capture the patterns of healthy brains aging. Then a stacked in ensemble to make a prediction on the test set (Al Zoubi et al., 2018). Deep neural networks and convolutional networks are used

TABLE 2.2: A summary of related work for predicting age from brain imaging data

Work	Data	No. of Samples	Performance
(Franke et al., 2010)	MRI	650	r = 0.92, MAE = 5 years
(Cole et al., 2018)	MRI	2,001	r = 0.96, MAE = 4.16 years
(AL, 2010)	fMRI	238	R^2 = 0.55
(Qin et al., 2015)	fMRI	183	MAE = 4.6 years
(Valizadeh et al., 2017)	MRI	3,144	R^2 = 0.77
(Dimitriadis and Salis, 2017)	EEG	94	R^2 = 0.6 for eyes open R^2 = 0.48 for eyes closed
(Liem et al., 2017)	fMRI + MRI	2,354	MAE = 4.29 years
(Al Zoubi et al., 2018)	EEG	468	R^2 = 0.37, MAE = 6.87 years, RMSE = 8.46 years
(Sun et al., 2019)	EEG of sleep	4,506	r = 0.83, MAE = 7.6 years
(Hogan et al., 2021)	subsequent EEG of sleep	86	Standard deviation for 2 nights = 4.7 years, 4 nights = 3 years

when researchers want to extract insights directly from neuroimaging data. The training task might be not only to extract patient brain age (Al Zoubi et al., 2018), but also sex (Van Putten, Olbrich, and Arns, 2018) and understand body movements (Schirrmeister et al., 2017).

2.6 State-of-the-art

It is hard to define any State-of-the-art solution, since there are different datasets and no single well-defined task, on which we can compare different approaches (as it is in machine translation or image recognition). As an evaluation metric most of the research use:

- R^2 - coefficient of determination, or r - correlation coefficient between actual subject age and estimated brain age;

- mean absolute error (MAE) or residual mean squared error (RMSE) as a difference between estimated brain age and actual age.

However, even if both research use R^2, it does not mean they are simply comparable. For example, (Dimitriadis and Salis, 2017) report achieved R^2= 0.6, while (Al Zoubi et al., 2018) get R^2= 0.4. However, they explicitly stated that results are not comparable since they define R^2 in a different way. We can understand the current SOTA by looking at the performance of various approaches and datasets 2.2.

As we may conclude from the table above, MRI-based methods overall perform better than EEG based. However, EEG researchers are getting better results last years, employing various approaches to measurements and analysis (EEG of sleep, subsequent EEG of sleep measurements).

Chapter 3

Approach to solution

3.1 Problem setting

We see the main research gap in the question: how to effectively feed EEG signal into machine learning algorithms? This includes selecting proper preprocessing steps and feature extraction pipeline. Currently, choosing feature extraction method heavily depends on the applications, and turns out to be a compromise between interpretation and performance.

To summarize, the main drawbacks of the current approaches are the following:

In preprocessing:

- manual EEG cleaning is time-consuming and requires high-paid qualified employees and routine work. It is also subjective and thus impossible to reproduce by other researchers;

- on the other hand, leaving artifacts as-is, without cleaning, may cause problems for feature extraction methods such as calculated features, that utilize statistical measures of the signal.

In feature extraction:

- **calculated features** are not fully interpretable and noise-sensitive;

- **images** also require defining many hyperparameters in advance, such as length of the frame and numbers of frames in the sequence of CNN input; same as the previous method, it also relies on existing domain knowledge, which may prevent us from finding new insights/patterns in the raw data;

- **raw signal** values are currently used mostly as input to deep neural networks; same as with CNN for images, their activation maps are hardly interpretable.

Thus, we will focus our later work on finding an approach for preprocessing and feature extraction, which allows us to overcome most of these gaps, thus has the following characteristics:

1. robust to artifacts, noise and requires minimal automatic preprocessing/cleaning;

2. extract features from signal values and discovers patterns of various lengths;

3. produce interpretable results, so it will be possible to trace what exactly in raw data has influenced a prediction.
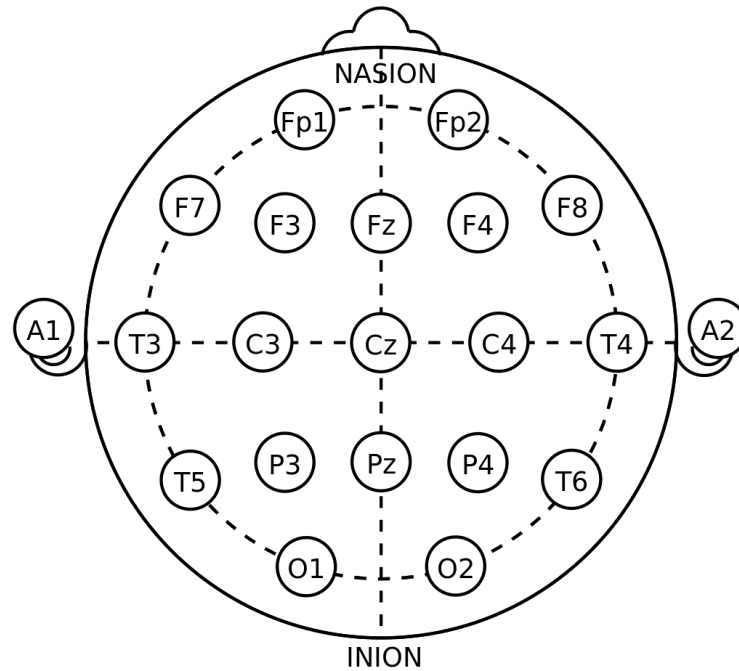
FIGURE 3.1: Electrode locations of International 10-20 system for EEG recording. Schematic top view of a scalp. Credits: Wikipedia

3.2 Dataset description

We analyzed routine clinical EEG recorded and evaluated in the process neurological assessment of patients in a public hospital in the Greater Vancouver area. The data is clinical, so it was not captured specifically for the experiment. The original sample included virtually all EEG scans (n=7048) recorded between 2012 and 2018 in the same EEG Unit, with several EEG stations, in general, by different EEG technicians.The hardware and firmware were identical across all the EEG stations, each equipped with a Natus Xltek EEG32U EEG amplifier (Natus, n.d.). The EEG montage was uniform as well: 10/20 system positioning, 22 EEG electrodes, two electrooculographic (EOG), and two electrocardiographic (ECG) electrodes 3.1. We take into account only signals from 20 channels - electrodes placed directly on a scalp. The location of the reference and ground electrodes was unknown. Each sample is 10-20 minutes of recorded EEG of a human brain, sampled with a frequency rate of 500 Hz or 512 Hz.

The data is a mix of inpatients and outpatients recordings. Inpatients are those who stay in a hospital for a long time, probably with serious pathology; outpatients who have an appointment in a hospital or clinic but do not need to stay overnight. Also, the data is a mix of male and female patients. We don't take into account conditions that patients have or the reason why they do EEG scans. We know the age of a patient for each sample. The age has a close to normal distribution from 0 to 100 years 3.2.

EEG data were converted from the Natus proprietary format into the EDF format with Natus Neurworks EEG software. If EEG recording in an original EEG study were turned off and then turned off again, potentially several times, the recorded EEG segments were linked with zeros in the EDF file. EEGs were deidentified with the PyEDFlib Python toolbox (Nahrstaedt, 2015-2022).
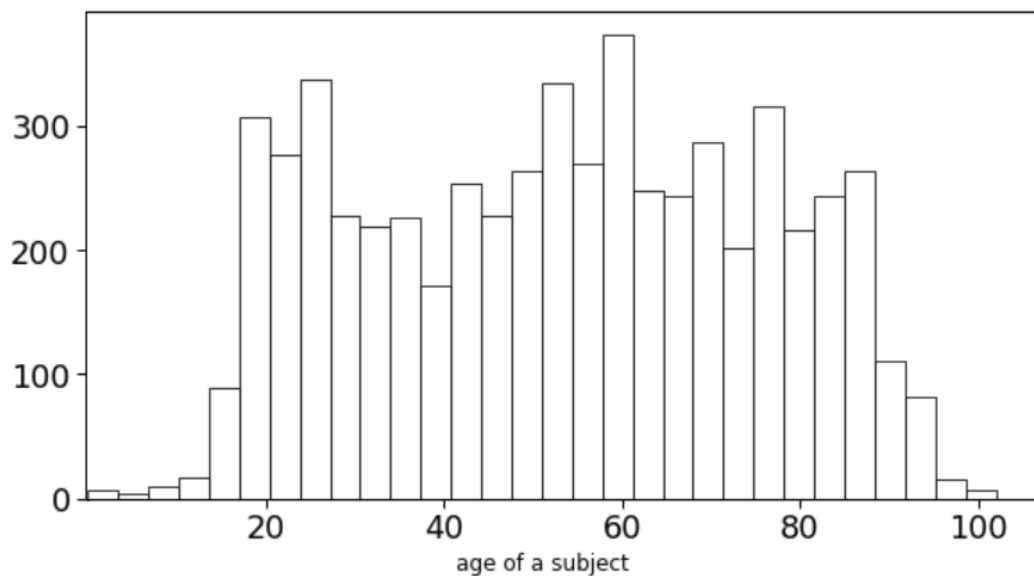
FIGURE 3.2: Distribution of EEG samples number across ages in the dataset

Before applying the pipeline described later, we have applied the following cleaning. We resample each recording's signal to 500 Hz frequency, since some recordings might have different sampling frequencies. Also the recordings contain intervals with no signal. It is th results of turned off equipment or disconnected electrode. So we have to remove these flat intervals without a signal (with zero signal).

We have also applied frequency filtering that keeps only signal with frequencies between 1 and 55 Hz, to exclude unwanted noise such as electricity grid frequency (it is 60 Hz in Canada) or sudden patients' moves, which have frequencies up to 1 Hz. In this way the signal still include major brain waves - from delta waves (0.5-4 Hz) to gamma (>35 Hz) (Abhang, Gawali, and Mehrotra, 2016). Also applied detrending to remove linear trend along time axis from data.

We have also removed intervals of special procedures performed on patients during recordings, such as hyperventilation (deep breathing) and photic stimulation (flashing light). Physicians apply these tests to patients in order to detect abnormal activity of a brain for epilepsy diagnosis. Since these procedures burst abnormal activity, and weren't performed for all subjects, we exclude them from the analysis. Finally we acquired 10 minutes clean intervals from each EEG, without flat intervals, hyperventilation and photic stimulation. Those recording that do not have a clean intervals of needed length were not included into final dataset. After this cleaning the new sample size is n=5850

3.3   Method

To identify variable length patterns in time series, we use the Byte Pair Encoding (BPE) compression technique (Gage, 1994). BPE has been around for a long time, however, since it was used in (Sennrich, Haddow, and Birch, 2015) it received much more recognition. BPE is a compression technique in which the most common pair of consecutive symbols is replaced by a new symbol. In (Sennrich, Haddow, and Birch, 2015) it was used to address the rare word problem in neural machine translation by
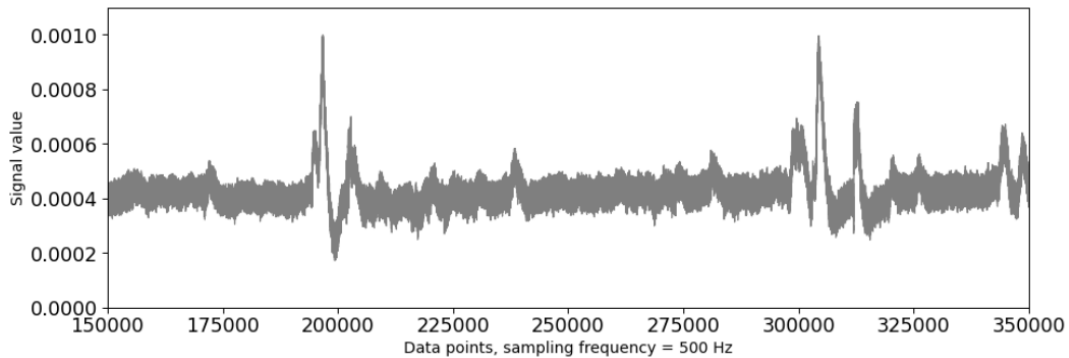
FIGURE 3.3: A slice of raw EEG signal, one of 20 channels, original
length = 200,000 data points

subword tokenization. Traditionally in NLP, lemmatized or stemmed words were considered tokens in text. With subword tokenization, now characters are considered as tokens and with each iteration, a pair of tokens that most frequently occurred together are merged together to build a new token. In this way, compound words like "authorship" could be understood by the model without having observed it beforehand and by breaking it into subwords "author-" and "-ship".

In the work (Tavabi and Lerman, 2021) authors have recently introduced the application of this approach to time series analysis. They have employed BPE to classify time series data of heart rate and step count from wearable devices to predict subjects' personality traits, which is close to our task of predicting patients' age based on EEG data. Thus, we reproduce their transformation approach in order to: (1) convert signal value time series into symbolic (character) series, (2) tokenize string, (3) use tokenized data to train a simple machine learning model for patients' age prediction.

3.3.1 Piecewise Aggregate Approximation (PAA)

PAA reduces the dimension of the input time series 3.3 by splitting them into segments of a given length (window size) and averaging the signal values in these segments (Keogh et al., 2001). As the result, it reduces input length and at the same time preserves patterns. Taking into account frequency sampling of 500 Hz, we want to keep the signal from gamma brain waves, which correspond to frequencies up to 50 Hz. Thus we use a window size of length 10 to average the signal 3.4.

3.3.2 Discretization / symbolization

We define outliers as data points outside +- 1.5 Inter Quartile Range (IQR). After setting the outliers aside, values are binned (discretized) by equal-width bins. In the original work (Tavabi and Lerman, 2021), both outlier detection and the binning are based on the entire dataset, as opposed to each time series independently. Defining outliers and binning the values based on the entire data helps better capture the differences between series. However, due to the nature of our data and the high variance of signal value across samples, we define outliers for each channel within each sample (EEG record). Each discrete bin is assigned a symbol (character), which we use to identify patterns.
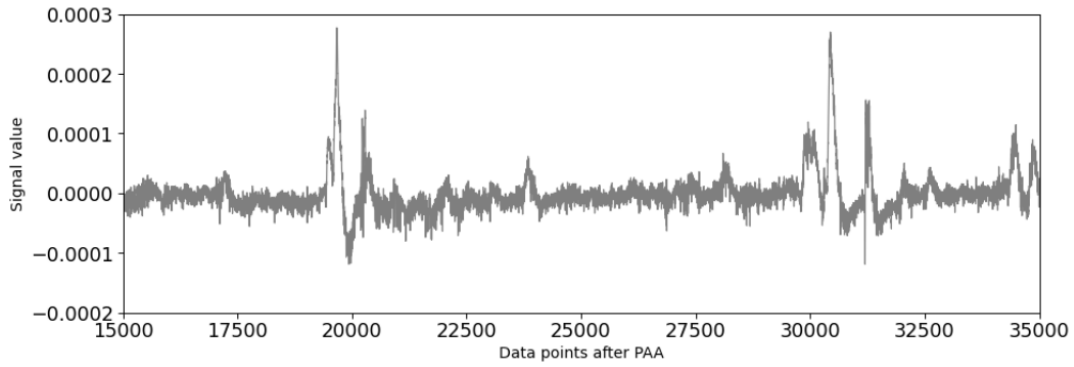
FIGURE 3.4: Same data as in 3.3 after PAA transformation with window size=10, resulted length = 20,000 data points
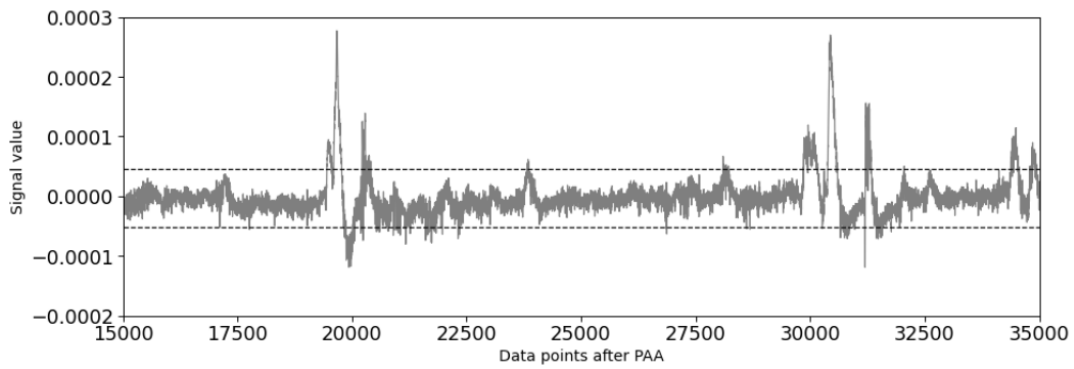

FIGURE 3.5: Slice of PAA transformed signal, with outliers boundaries (dashed)
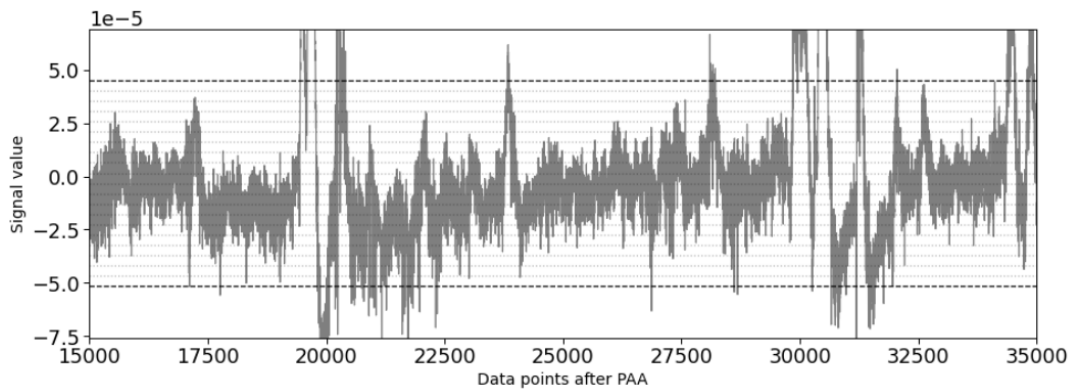

FIGURE 3.6: Same data 3.5 'zoomed in' between outlier boundaries, with dotted lines indicating discretization bins.

Input string: **D C B B A B C D C B A B C D**

Step 1: most frequent is pair of B and A, merge B+A:
D C B BA B C D C BA B C D

Step 2: most frequent is pair of BA and B, merge BA + B
D C B BAB C D C BAB C D

Step 3: most frequent is pair of BAB and C, merge BAB + C
D C B BABC D C BABC D

Step 4: most frequent is pair of D and C, merge D + C
DC B BABC DC BABC D

Final vocabulary: **B, D, DC, BABC**

FIGURE 3.7: Example of BPE on discretized time series: The first line
shows the original discretized series. In each iteration, the most com-
mon pair is identified and joined in a new token.

The output of such discretization is a string of characters, each representing a bin
that the signal value belongs: *ijjijiiijklllllllhfgilligjfefhighillljjlihhjjijjjklllllljhjheffgihlji...*

3.3.3 Patterns identification

Here we applied BPE algorithm, illustrated in Figure 4. Another difference from
the original paper, here we apply the existing Hugging Face implementation of BPE
algorithm (HuggingFace, 2021). BPE creates a base vocabulary consisting of all sym-
bols that occur in the set of unique words and learns merge rules to form a new
symbol from two symbols of the base vocabulary 3.7. It does so until the vocabulary
has attained the desired vocabulary size. We train tokenizer on the whole corpus of
symbolic series that we have in the dataset.

3.3.4 Tokens to features

Then we tokenize each treat each token as a 'word' and count the number appear-
ances of them for each sample, similar to the bag-of-words approach, divided by the
length of sample, weighted by the length of the token. Thus we get the frequency of
each token appearance in the sample.

$$\text{Token Frequency}, \% = \frac{\text{No. of appearances} \times \text{No. of the symbols in the token}}{\text{Total number of symbols in the sample}} \times 100$$

We have applied tokenizer separately to each channel, occurence frequency also cal-
culated for tokens per channel 3.8. Thus resulted dimensionality is approximately

	Channel 1 (C4)			Channel 2 (C3)			...
	token 'DC'	token 'BABC'	...	token 'DC'	token 'BABC'	...	...
Sample X	1.5	0.05	...	0.001	0.35	...	...

Sum up to 100   Sum up to 100

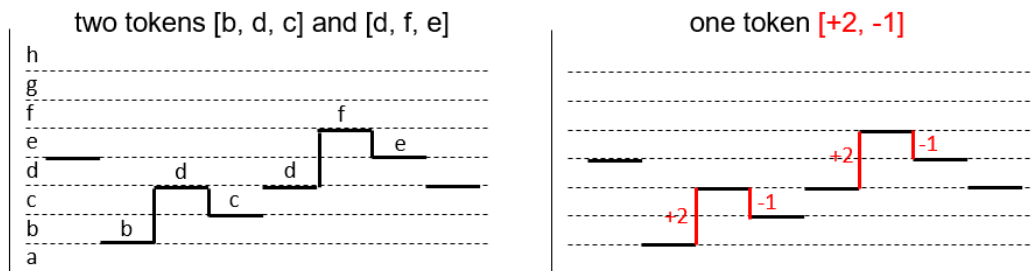FIGURE 3.8: Scheme of preprocessed and tokenized dataset later used for age prediction

FIGURE 3.9: Example of treating symbolic tokens with an absolute or relative position

equal to VS multiplied by number of channels, where VS is vocabolary size of tokenizer. 'Approximately' is because some tokens might not appear across all samples in a specific channel, so the channel lacks the corresponding token feature.

3.3.5 Relative tokens

Our goal is to identify patterns that carry some information relevant for, in our case, the age prediction task. Thus we should take into account cases when two tokens have the same shape but have different absolute positions. In order to handle these cases, we transform tokens from absolute to relative form, calculating the distance between symbols. Here is the example 3.9: for token 'bdc', the distance between 'b' and 'd' is 2 bins up, and between 'd' and 'c' is 1 bin down. Thus the token now having the form [+2, -1]. Same with the token 'dfe'. In this way instead of having two different tokens/features - we have only one.

Later we used preprocessed data both with absolute and relative tokens for age prediction in order to compare performance.

3.3.6 Train model for age prediction

Since the work is focuses more on feature extraction stage for EEG classification, we pick the simplest model that can capture non-linear relations. Here we apply the Random Forest Regressor model, where features are frequencies of tokens/patterns appearances, and the target label is patients' age. The dataset was split into train and test subsets (80% train and 20% test) with a constant random seed. We use cuML RAPIDS implementation of the Random Forest model (RAPIDS, n.d.) to utilize high-performing training on GPU (Tesla T4 16GB).

3.3.7   Hyperparameters tuning

The whole pipeline has the following hyperparameters:

- WS - window size for PAA transformation, $10 < WS < 200$

- NB - number of bins (characters) for symbolization, $10 < NB < 20$

- MF - minimal frequency of the token appearance in the corpus, required for tokenizer to include the token into the vocabulary. We manually fit MF to obtain a reasonable vocabulary size of 1500 tokens per channel.

- NE - number of estimators for Random Forest Regressor age prediction, $100 < NE < 2000$

- MD - max depth of each tree for Random Forest Regressor age prediction, $8 < MD < 32$

3.3.8   Evaluation

We evaluate the performance of the age prediction model with the following metrics:

- **Mean Absolute Error (MAE)** between actual and predicted subject's age in years. Our goal is to minimize it.

- **Pearson correlation coefficient and distance correlation coefficient**. Distance correlation also takes into account non-linear relations in the data (Richards, 2017).

- **Explained variance** score which measures the proportion to which a model accounts for the variation of a dataset.

Chapter 4

Results

4.1 Age prediction

In a course of the project we have run multiple experiments. We were continuously increasing the number of samples in the dataset and the level of detalization of our symbolization pipeline. The issue with symbolization parameter tuning is that it's impossible to fix some hyper-parameters and change others. When we change window size of PAA transformation, we change the resolution and thus - the number of tokens in the dataset. Changing the number of bins within the discretization step, we change the number of unique symbols. This also affects the number of unique tokens in the tokenizer vocabulary. That is why we can not fix the minimal frequency of tokenizer: changing any of the previous parameters completely change number of tokens in our corpus and requires new value of minimal frequency, at least to maintain the same vocabulary size. This fact makes hyper-parameter tuning more complicated and required manual adjustments.

We have run first experiments employing only one of 20 channels (C3) form each EEG recording. Then we switched to using all channels. Also due to high computing time for Random Forest training on CPU's, for later experiments we switched into training the model on GPU with RAPIDS framework. The progress of improving MAE score and experiment setup is described in the following table 4.1

TABLE 4.1: Experiments setup and corresponding MAE score in years

Ex #	No of samples	WS	NB	MF	RF Model	NE	MD	MAE
1	1237 (C3 only)	200	10	2	sklearn	100	default	**21.3**
2	2287 (C3 only)	200	10	10	sklearn	100	default	**18.0**
3	4380 (C3 only)	200	10	200	sklearn	100	default	**17.6**
4	4380 (C3 only)	200	20	100	sklearn	100	default	**17.5**
5	5850 (C3 only)	200	20	200	sklearn	100	default	**17.7**
6	5850 (C3 only)	200	20	200	sklearn	500	default	**17.7**
7	5850 (C3 only)	200	20	1000	sklearn	100	default	**17.5**
8	5850 (C3 only)	200	20	5000	sklearn	100	default	**17.8**
9	5850 (all chs.)	200	20	20000	sklearn	100	default	**17.2**
10	5850 (all chs.)	200	20	20000	RAPIDS	1000	16	**16.9**
11	5850 (all chs.)	50	20	80000	RAPIDS	1000	16	**16.3**
12	5850 (all chs.)	10	20	120000	RAPIDS	1000	16	**15.9**
13	5850 (all chs.)	10	20	120000	RAPIDS	1000	32	**15.9**
14	5850 (all chs., relative tokens)	10	20	120000	RAPIDS	1000	32	**15.9**

The best performing model was trained on the tokens made of symbolic series obtained by transforming signal with window size = 10 and number of bins = 20. Resulting dataset has a size of 5850 samples with 32,000 features, representing the appearance frequency of a specific token in a sample.

The model has following parameters n_estimators = 1000, max_depth = 16. Other parameters were not included in tuning, since we were focused more on preprocessing steps and applying BPE and accurate prediction of the age itself is outside of the scope of this work.

The model has predicted subjects' age with

- MAE = 15.9 years,

- Pearson correlation coefficient = 0.56,

- distance correlation coefficient = 0.54,

- and explained variance = 0.287

We also have transformed the dataset of token frequencies from absolute symbolic tokens, of a form 'acb', to relative change tokens, of a form (+1. -1), and trained a new model with the same parameters tuned 4.1. The new model hasn't shown a significant improvement in performance:

- MAE = 15.9 years,

- Pearson correlation coefficient = 0.54,

- distance correlation coefficient = 0.57,

- and explained variance = 0.294

However, this transformation has merged multiple instances of the same relative tokens so resulted dataset has 8,000 features instead of 32,000. This has significantly improved the computational time required for the model training: 40 minutes versus 90 minutes.

4.2 Interpretability and traceability

The proposed method of preprocessing and feature extraction with BEP ensures full interpretability. We know the exact positions of the token and can trace it in the original EEG signal. Since our input signal has a frequency of 500 Hz and we averaged it with window size 10 data points, the token of two symbols has a duration of just 0.04 seconds. The method allows finding the corresponding fragments for the most influential tokens, that correlate the most with age 4.2

4.3 Tokens importance

Besides modeling for age prediction, we explore how token frequencies are changing with age. Here we also employed the distance correlation coefficient in order to capture nonlinear dependencies, and calculate it for each token. Top correlated tokens have distance correlation coefficient values from 0.25 to 0.28. We check how frequencies of these most correlated token changes with age and find out clear nonlinear dependency. The results are similar for relative tokens as well 4.3
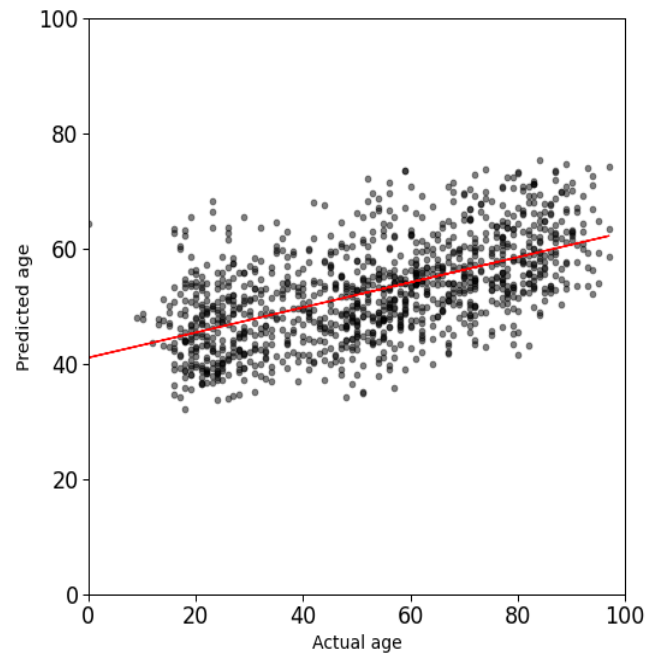
FIGURE 4.1: Visualised relations between actual and predicted age on the dataset with relative tokens. Each dot represents a sample from the test subset, with a fitted linear regression line. For ideal model predictions, the line should be a bisector.

During the analysis of tokens correlation, we observe that it varies across EEG channels. Tokens build based on the EEG signal from electrodes placed on frontal and occipital areas of the scalp have a higher median and mean distance correlation with age 4.4. It also preserves spatial symmetry between the left and right parts of the scalp, which was not initially indicated in the dataset. Channel was marked only as text label: 'O2', 'Fpz', etc.
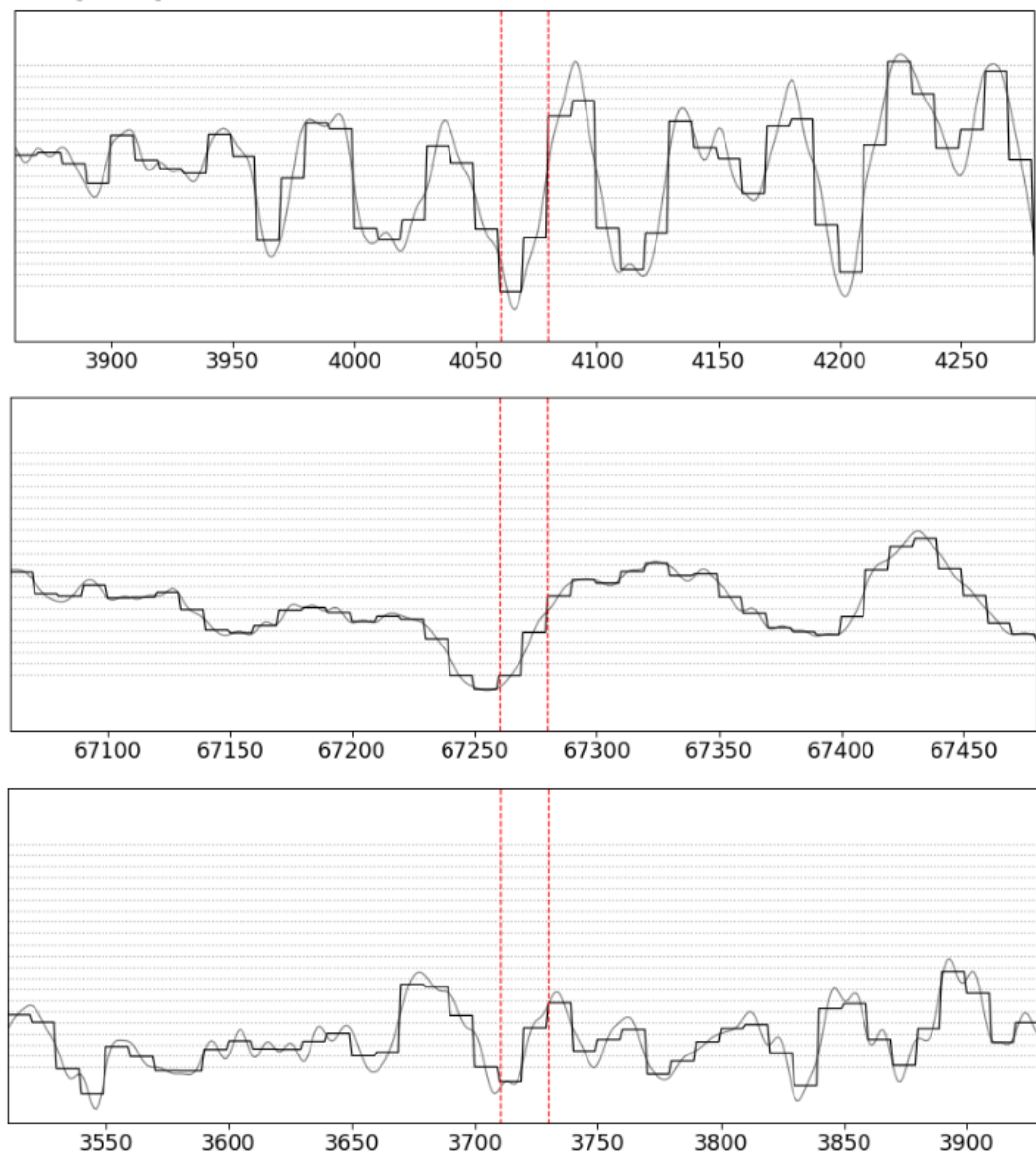
FIGURE 4.2: Highlighted parts of the raw signal that correspond to the token 'ae' from channel O2 in different samples. Original EEG signal in a light-gray and PAA-transformed (averaged) in black. Each plot has a duration of 0.84 seconds, the token duration is 0.04 seconds.
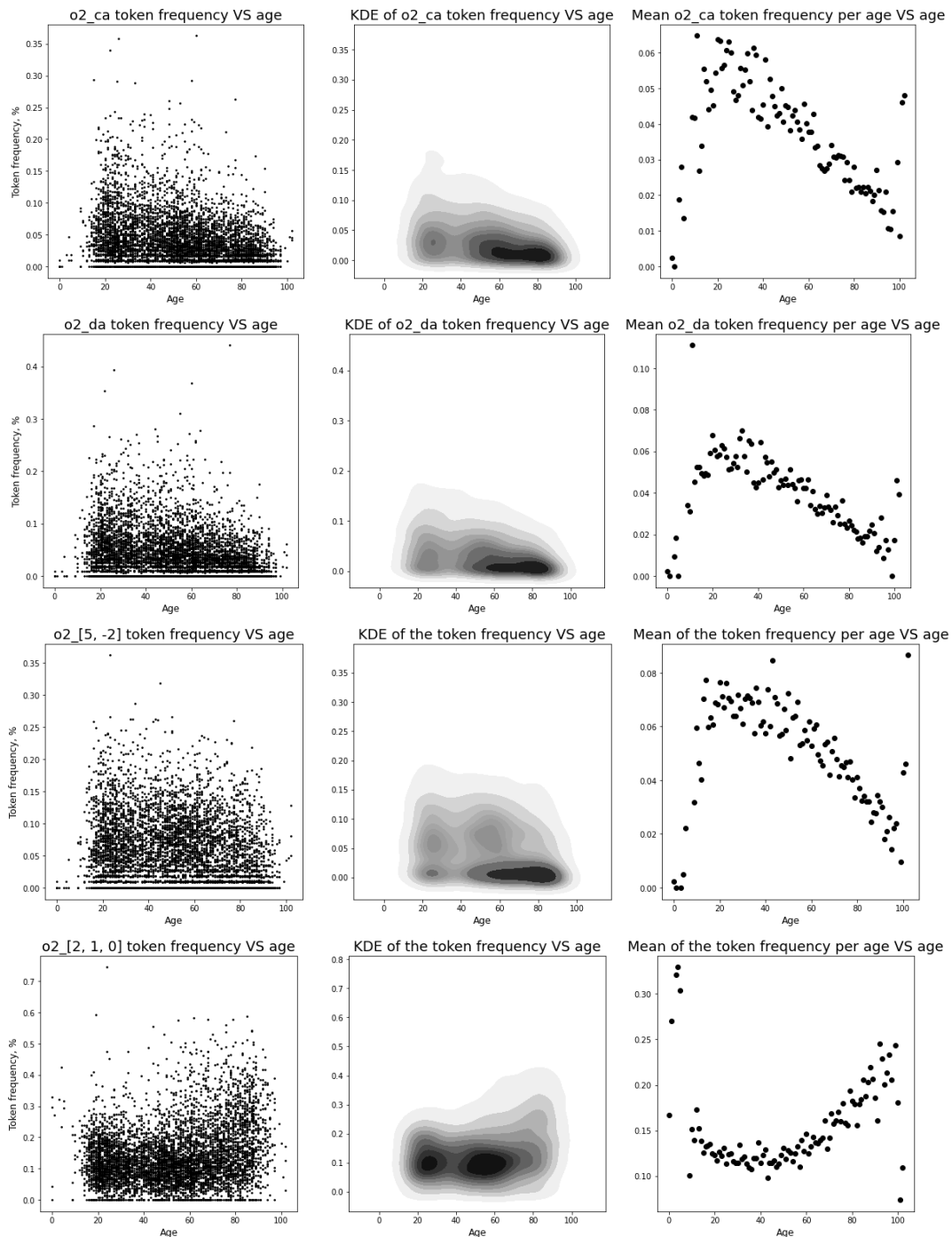
FIGURE 4.3: Relation of token appearance frequency in the sample with age. There are two symbolic tokens 'ae' and 'ea' and two relative [+5, -2] and [+2, +1, 0] from EEG channel O2. The first subplot shows a token frequency for each sample corresponding to age. The second one is the same but visualized in a form of 2D Kernel Density Estimation: darker areas correspond to more dense observation placement. The third subplot shows a mean token frequency grouped for each year from 0 to 100. The mean frequency of some tokens declines with age, while others - increase.

FIGURE 4.4: Schematic top view of a subject's head. Each cell represents a specific EEG channel - an area of the scalp where an electrode sampling the signal. 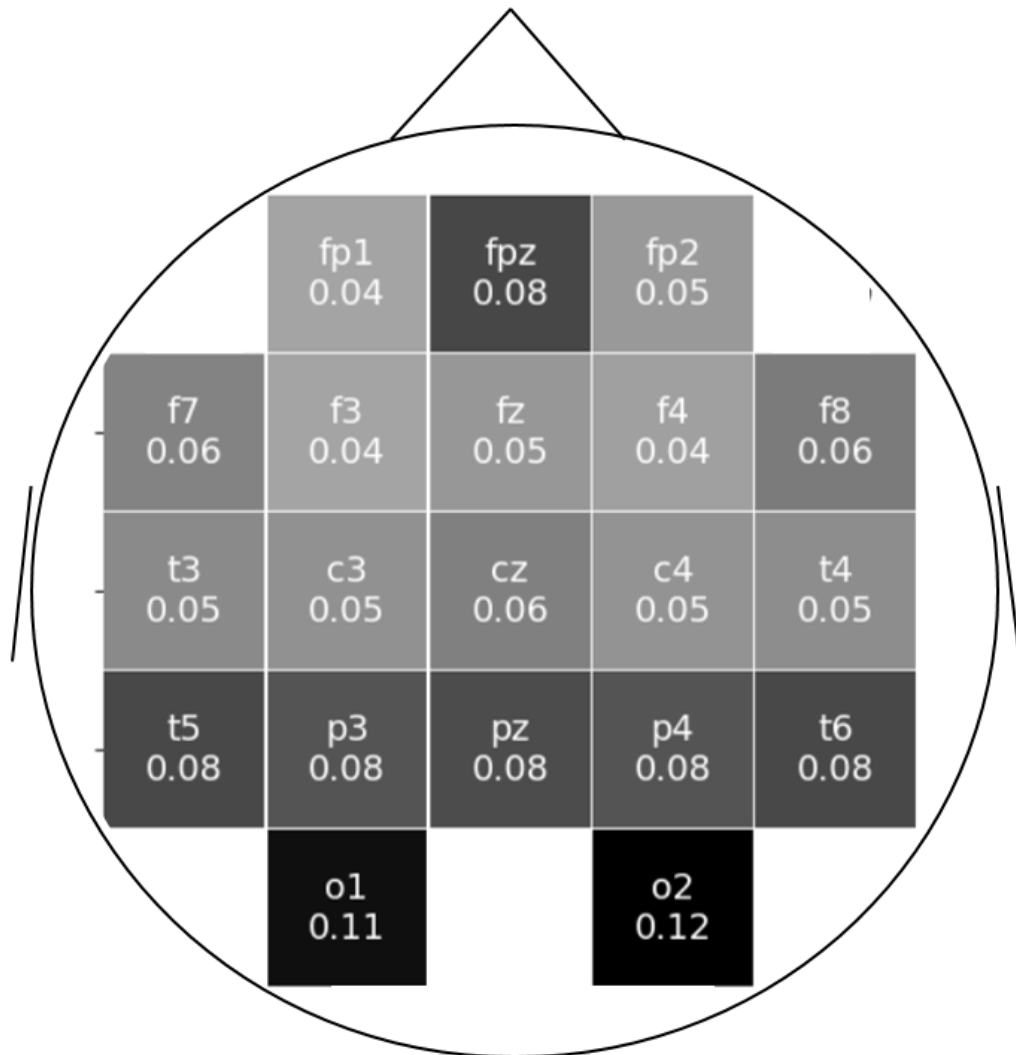The labels indicate a name of EEG channel with the median distance correlation of its tokens frequencies with subjects' age

Chapter 5

Discussion

We have introduced an automatic method of preprocessing EEG signals with time-series symbolization and feature extraction with tokenization based on the Byte-Pair Encoding (BPE) algorithm. The algorithm can handle EEG recordings of arbitrary length. It is also able to discover important task-related patterns of a variable duration starting from 0.05 seconds, with no need to define the intended pattern length in advance.

The features constructed with our method enable us to capture the spatial dependency between channels and the left-right symmetry of the brain. This observation proves that the method is able to gain functional information about the brain which allows for making reasonable inferences.

The accuracy of a subject's age prediction measured with MAE is quite low compared to other works in the field of age prediction based on EEG 5.1

TABLE 5.1: Comparison of performance with best EEG based age prediction algorithms

Source	Data	of samples	Performance
(Al Zoubi et al., 2018)	EEG	468	MAE = 6.87 years
(Sun et al., 2019)	EEG of sleep	4506	MAE = 7.6 years
our work	clinical EEG	5850	MAE = 15.9 years

However, we should take into account that due to the nature of the dataset the results are barely comparable. Our dataset contains clinical EEG recordings that were not collected for research purposes. The recordings were taken over a long period of time in different hospitals, by different technicians. We also do not differentiate samples by outpatients and inpatients, some of them might have severe conditions. We focus our later work on applying other approaches for EEG feature extraction to the same dataset to have a benchmark for age prediction performance.

The feature extraction method we demonstrated can serve for EEG analysis applying to other tasks outside age prediction. The algorithm can highlight specific regions of interest in EEG depending on the task. This traceability of features to specific fragments in EEG opens possibilities for using it in decision-support systems for neurophysiologists who manually assess EEG recordings.

Bibliography

Abhang, Priyanka A, Bharti Gawali, and Suresh Mehrotra (2016). *Introduction to EEG-and speech-based emotion recognition*. Academic Press.

AL, Dosenbach NUF Nardos B Cohen (2010). "Fair DA Power JD Church JA Nelson SM Wig GS Vogel AC Lessov-Schlaggar CN Barnes KA Dubis JW Feczko E Coalson RS Pruett JR Barch DM Petersen SE Schlaggar BL 2010 Prediction of individual brain maturity using fMRI Science 32913581361". In: *Dosenbach NUF, Nardos B, Cohen AL, Fair DA, Power JD, Church JA, Nelson SM, Wig GS, Vogel AC, Lessov-Schlaggar CN, Barnes KA, Dubis JW, Feczko E, Coalson RS, Pruett JR, Barch DM, Petersen SE, Schlaggar BL*, pp. 1358–1361.

Al Zoubi, Obada et al. (2018). "Predicting age from brain EEG signals—A machine learning approach". In: *Frontiers in aging neuroscience* 10, p. 184.

Butler, Ellyn R et al. (2020). "Statistical pitfalls in brain age analyses". In: *bioRxiv*.

Cole, James H et al. (2018). "Brain age predicts mortality". In: *Molecular psychiatry* 23.5, pp. 1385–1392.

Craik, Alexander, Yongtian He, and Jose L Contreras-Vidal (2019). "Deep learning for electroencephalogram (EEG) classification tasks: a review". In: *Journal of neural engineering* 16.3, p. 031001.

Dimitriadis, Stavros I and Christos I Salis (2017). "Mining time-resolved functional brain graphs to an EEG-based chronnectomic brain aged index (CBAI)". In: *Frontiers in human neuroscience* 11, p. 423.

Farnsworth, Bryn (2019). *EEG vs. MRI vs. fMRI – What are the Differences?* `https://imotions.com/blog/eeg-vs-mri-vs-fmri-differences/`. [Online; accessed 20-June-2022].

Franke, Katja and Christian Gaser (2019). "Ten years of BrainAGE as a neuroimaging biomarker of brain aging: what insights have we gained?" In: *Frontiers in neurology*, p. 789.

Franke, Katja et al. (2010). "Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: exploring the influence of various parameters". In: *Neuroimage* 50.3, pp. 883–892.

Gage, Philip (1994). "A new algorithm for data compression". In: *C Users Journal* 12.2, pp. 23–38.

Hennig, Jürgen et al. (2003). "Functional magnetic resonance imaging: a review of methodological aspects and clinical applications". In: *Journal of Magnetic Resonance Imaging: An Official Journal of the International Society for Magnetic Resonance in Medicine* 18.1, pp. 1–15.

Hillman, Elizabeth MC (2014). "Coupling mechanism and significance of the BOLD signal: a status report". In: *Annual review of neuroscience* 37, p. 161.

Hogan, Jacob et al. (2021). "Night-to-night variability of sleep electroencephalography-based brain age measurements". In: *Clinical Neurophysiology* 132.1, pp. 1–12.

HuggingFace (2021). *Byte-Pair Encoding tokenization*. `https://huggingface.co/course/chapter6/6`. [Online; accessed 4-June-2022].

Jung, Bernd André and Matthias Weigel (2013). "Spin echo magnetic resonance imaging". In: *Journal of Magnetic Resonance Imaging* 37.4, pp. 805–817.

Keogh, Eamonn et al. (2001). "Dimensionality reduction for fast similarity search in large time series databases". In: *Knowledge and information Systems* 3.3, pp. 263–286.

Liem, Franziskus et al. (2017). "Predicting brain-age from multimodal imaging data captures cognitive impairment". In: *Neuroimage* 148, pp. 179–188.

Mills, Andrew F et al. (2017). "Principles of quantitative MR imaging with illustrated review of applicable modular pulse diagrams". In: *RadioGraphics* 37.7, pp. 2083–2105.

Nagel, Sebastian (2019). "Towards a home-use BCI: fast asynchronous control and robust non-control state detection". PhD thesis. Universität Tübingen.

Nahrstaedt, Holger (2015-2022). *PyEDFlib -EDF/BDF Toolbox in Python*. https://pyedflib.readthedocs.io/. [Online; accessed 4-June-2022].

Natus (n.d.). *Natus Xltek EEG32U EEG amplifier*. https://neuro.natus.com/products-services/natus-eeg32u-eeg-amplifier. [Online; accessed 4-June-2022].

Noachtar, Soheyl and Jan Rémi (2009). "The role of EEG in epilepsy: a critical review". In: *Epilepsy & Behavior* 15.1, pp. 22–33.

Puvill, Thomas et al. (2016). "Impact of physical and mental health on life satisfaction in old age: a population based observational study". In: *BMC geriatrics* 16.1, pp. 1–9.

Qin, Jian et al. (2015). "Predicting individual brain maturity using dynamic functional connectivity". In: *Frontiers in human neuroscience* 9, p. 418.

RAPIDS (n.d.). *RAPIDS. Open GPU Data Science*. https://docs.rapids.ai/api/cuml/stable/api.html. [Online; accessed 4-June-2022].

Richards, Donald St P (2017). "Distance correlation: A new tool for detecting association and measuring correlation between data sets". In: *arXiv preprint arXiv:1709.06400*.

Schirrmeister, Robin Tibor et al. (2017). "Deep learning with convolutional neural networks for EEG decoding and visualization". In: *Human brain mapping* 38.11, pp. 5391–5420.

Sennrich, Rico, Barry Haddow, and Alexandra Birch (2015). "Neural machine translation of rare words with subword units". In: *arXiv preprint arXiv:1508.07909*.

Smith, Stephen M et al. (2019). "Estimation of brain age delta from brain imaging". In: *Neuroimage* 200, pp. 528–539.

Sun, Haoqi et al. (2019). "Brain age from the electroencephalogram of sleep". In: *Neurobiology of aging* 74, pp. 112–120.

Tavabi, Nazgol and Kristina Lerman (2021). "Pattern Discovery in Time Series with Byte Pair Encoding". In: *arXiv preprint arXiv:2106.00614*.

Teplan, Michal et al. (2002). "Fundamentals of EEG measurement". In: *Measurement science review* 2.2, pp. 1–11.

Toole, John MO' and Geraldine B Boylan (2017). "NEURAL: quantitative features for newborn EEG using Matlab". In: *arXiv preprint arXiv:1704.05694*.

Valizadeh, SA et al. (2017). "Age prediction on the basis of brain anatomical measures". In: *Human brain mapping* 38.2, pp. 997–1008.

Van Putten, Michel JAM, Sebastian Olbrich, and Martijn Arns (2018). "Predicting sex from brain rhythms with deep learning". In: *Scientific reports* 8.1, pp. 1–7.