# Byte-Pair Encoding for classifying routine clinical electroencephalograms

Mykola Klymenko, Sam M. Doesburg, George Medvedev, Pengcheng Xi,
Urs Ribary, Vasily A. Vakorin

***Abstract*— Routine clinical EEG is a standard test used for the neurological evaluation of patients. A trained specialist interprets EEG recordings and classifies them into clinical categories. Given time demands and high inter-reader variability, there is an opportunity to facilitate the evaluation process by providing decision support tools that can classify EEG recordings automatically. The classification models are expected to be interpretable, able to process EEG of varying durations, and recorded in heterogeneous imaging environments. Our study aimed to test and validate a framework for EEG classification which satisfies these requirements by considering EEG as unstructured text. The framework was thus based on symbolizing EEG signals and adapting a previously proposed method from natural language processing (NLP). We considered a highly heterogeneous and extensive sample of routine clinical EEGs (n=5850), with a wide range of participants aged between 15 and 99 years. We symbolized the multichannel EEG time 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 a framework, we used newly-reconstructed EEG features to predict patients' biological age using a classical (non-neural) machine learning model. We also correlated tokens' occurrence frequencies with age. We found that the age prediction model achieved a mean absolute error of 15.6 years. The highest correlations between the frequencies of tokens and age were observed at frontal and occipital EEG channels. Our findings demonstrated the feasibility of applying an NLP-based approach to classifying clinical EEG. Notably, the proposed algorithms could be instrumental in classifying clinical EEG with minimal preprocessing and identifying clinically-relevant short events, such as epileptic spikes.**

***Index Terms*— byte-pair encoding, machine learning, EEG classification, clinical electroencephalography, age prediction, natural language processing, multivariate time series.**

This work was supported in part by the National Research Council (NRC) of Canada, Collaborative Research and Development Grant DHGA-116-1, and the Digital Research Alliance of Canada, Research Platforms and Portals (RPP) grant.

M. Klymenko, Applied Science Faculty, Ukrainian Catholic University, Ukraine (e-mail: mykola.klymenko@ucu.edu.ua)

S. M. Doesburg, Department of Biomedical Physiology and Kinesiology, Behavioral and Cognitive Neuroscience Institute, Simon Fraser University, Canada (e-mail: sam_doesburg@sfu.ca)

G. Medvedev, Division of Neurology, Fraser Health Authority, Canada (email: gmedvedev@gmail.com)

X. Pengcheng, Digital Technologies Research Centre, National Research Council of Canada, Canada (email: pengcheng.xi@nrc-cnrc.gc.ca)

U. Ribary, Behavioral and Cognitive Neuroscience Institute, Department of Psychology, Simon Fraser University, Canada (email: urs_ribary@sfu.ca)

V. A. Vakorin, Department of Biomedical Physiology and Kinesiology, Behavioral and Cognitive Neuroscience Institute, Simon Fraser University, Canada (email: vasily_vakorin@sfu.ca)

## I. Introduction

EEG is a neurophysiological test that records the brain's electrical activity by measuring time-varying electrical potential differences between pairs of electrodes. In clinical practice, EEG is used to help diagnose several clinical conditions and symptoms. EEG is affordable and captured in a standardized fashion, making it widely available in hospitals throughout the world. The conventional approach to clinical EEG evaluation consists of an analysis of data visually presented on a computer screen by a highly trained expert. The expert has to describe clinically relevant waveforms and differentiate EEG records into broad categories, i.e., normal or abnormal, and if abnormal, epileptiform or non-epileptiform, according to the American Clinical Neurophysiology Society guidelines (ACNS, 2022). Such an approach to EEG evaluation has several challenges. Numerous features in the data are distributed over several channels, and these features vary over time. Evaluating EEG patterns may be affected by personal and institutional biases, which results in a high inter-interpreter variability (Schneider et al., 2003). For example, certain features of sleep EEG (Asadi-Pooya et al., 2019), common normal variants (Kang et al., 2019), or EEG artifacts (Mathias et al., 2019) can be misinterpreted as pathological discharges.

The problem of automatic classification of EEG signals using machine learning techniques has been increasingly studied in cognitive and clinical neuroscience. Potential applications include emotion recognition (Teplan, 2002), motor imagery tasks (Pfurtscheller and Neuper, 2001), seizure detection (Andrzejak et al., 2001), brain injury assessments (Vakorin et al., 2016), Alzheimer's classification (Kim and Kim, 2018), depression (Acharya et al., 2018), gender classification (Guo et al., 2015), and detection of abnormal EEG (Van Leeuwen et al., 2019), to name a few.

From a methodological perspective, several approaches exist through feeding EEG data into machine learning models. Studies on classifying EEG signals can be loosely divided into three categories based on how EEG is organized as an input for prediction models: time series, images, and a vector of

extracted features (Craik et al., 2019). First, the time series approach preserves the original dynamics recorded with EEG. In this case, the input is highly dimensional, which can be suitable for deep learning, but not for classical (non-neural) approaches. Second, feature extraction is often used to reduce the dimensionality of EEG signals. Extracted features can be of various natures. Often these features are defined in the frequency domain, i.e., spectral power, in the time-frequency domain, i.e., spectrograms, or in the temporal domain, wherein various linear and nonlinear features are computed (extracted), e.g., signal complexity measures (Stancin et al., 2021). Feature extraction often requires advanced EEG preprocessing, as artifacts may bias the estimation of EEG features. Feature extraction can be naturally combined with deep neural networks or classical machine learning models. The features are often organized as a vector without a clearly defined structure. Third, raw EEG or EEG features can be organized as images. This common approach allows one to adopt methods that have demonstrated high performance in classifying images with deep learning, i.e., convolutional neural networks (CNN) (Schirrmeister et al., 2017).

Classification of clinical EEG is expected to have extra challenges compared to EEG recordings in a controlled laboratory setting. Clinical imaging environments are highly heterogeneous, often involving multiple EEG systems operated by multiple EEG technicians who may follow different guidelines for EEG recordings. In particular, those guidelines do not necessarily impose strict standards for selecting and locating the reference and ground electrodes. The abundance of physiological artifacts and their variability are expected to be higher in patients compared to participants tested in a laboratory. Notably, the duration of clinical EEG recordings varies significantly.

The heterogeneous nature of clinical EEG can potentially be addressed with tools developed to analyze and interpret unstructured data such as text. Natural Language Processing (NLP) is a field defined at the intersection of linguistics and computer science. The NLP works with human language and analyzes large amounts of symbolic data. A large number of NLP studies have delivered models of high performance, e.g., in applications for modern machine translation and generating human-like text, known as GPT-3 (Brown et al., 2020). The common ground of NLP methods is based on breaking unstructured text into repeating parts, such as characters or words, parts of words, or groups of words – so-called tokens. Subsequently, these tokens can be used as new features. In particular, tokens' occurrence has been used as features for various text classification tasks, e.g., predicting whether a social media post is expressing positive or negative feelings (Wang et al., 2014).

Recently, there have been attempts to adapt NLP tools for time series classification. One study exploited the analogy between NLP text patterns and signal patterns for anomaly detection in multivariate time series arising from telemetry streams (Horak et al., 2022). Another study classified the dynamics of heart rate and daily step count data from wearable devices to predict participants' personality traits (Tavabi and Lerman, 2021). In particular, the authors performed several steps in their analysis: (1) converting the original time series into a symbolic (character) series, (2) defining repeating parts (tokens) in the newly reconstructed symbolic series, and finally, (3) using tokens' occurrence frequencies to train their machine learning models. Some elements of such an approach can be traced to EEG studies, which characterized EEG signals in terms of symbolic complexity measures (Hussain et al., 2021), (Jordan et al., 2011), (Liu et al., 2006).

We hypothesized that NLP algorithms designed for analyzing semi-structured or unstructured data (texts) can be incorporated into an EEG preprocessing workflow with a subsequent EEG classification. In our study, we aimed to apply, test, and validate an NLP-based pipeline previously developed for time series classification (Tavabi and Lerman, 2021). We analyzed a very large sample of routine clinical EEG recorded in a highly heterogeneous cohort of patients between 15 and 99 years old. Having applied a minimalistic preprocessing pipeline, we converted EEG signals into strings of symbols. We then applied an algorithm known as byte-pair encoding to split the newly reconstructed text into the most frequent combinations of symbols by iteratively counting the appearances of unique pairs of symbols and merging the most frequent pairs into new complex symbols (tokens). To demonstrate the performance of our approach, we tested to what degree the most frequent tokens, associated with specific patterns of changes in EEG amplitude, could predict patients' biological age.

## II. METHODS

### *A. Dataset description*

We analyzed routine clinical EEG recorded and evaluated in the process of neurological assessment of patients in a hospital in the greater Vancouver area within Fraser Health Authority. The ethics protocol was approved by the Research Ethics Boards at Simon Fraser University and Fraser Health Authority, April 01, 2022, protocol number H18-02728. The original sample included virtually all EEG studies (n=7048) recorded between 2012 and 2018. The duration of recordings varied from 10 minutes to several hours (mean duration ~35 minutes). The patients' age range was between 15 and 99 years. The hardware and firmware were identical across all the EEG stations, each equipped with a Natus Xltek EEG32U EEG amplifier. The EEG montage was kept uniform: 10/20 system positioning, 20 standard EEG electrodes (FP1, FPZ, FP2, F3, F4, F7, F8, FZ, T3, T4, T5, T6, C3, C4, CZ, P3, P4, PZ, O1, O2), two electrooculography (EOG), and two electrocardiographic (ECG) electrodes. The location of the reference and ground electrodes was unknown. The sampling frequency was either 500 Hz or 512 Hz.

### *B. EEG preprocessing*

We applied a minimal set of EEG preprocessing procedures. First, EEG data were converted from the Natus proprietary format into the EDF format with Natus's Neuroworks. If EEG recordings in an original EEG study were turned off and on, potentially several times, the recorded EEG segments were linked with digital zeros in the resulting continuous EDF file. EEGs were then de-identified with the PyEDFlib Python toolbox (Nahrstaedt, 2022). Each EEG recording was filtered between 0.5 and 55 Hz and then resampled to 500 Hz when the original sampling frequency was not 500 Hz. Separately for each EEG channel, we removed a possible linear trend. In each EEG scan, we identified the time intervals corresponding to the flat signal (digital zeros), hyperventilation, and photic stimulation procedures, if any. Avoiding these time intervals, we aimed to randomly select one 10-minute EEG segment from each of the original EEG recordings, which failed in some cases. These cases were discarded from further analysis. The final sample included n=5785 EEG segments, each associated with one original EEG scan.

### *C. Symbolization of EEG time series*

To apply NLP methods, we had to convert EEG signals into text. This procedure was divided into two stages: Piecewise Aggregate Approximation (PAA) and discretization. PAA was applied across time, whereas discretization was applied across EEG amplitude (Keogh et al., 2001). At the PAA stage, we divided the entire EEG segment into non-overlapping windows of a fixed length of 10 data points. The window length was chosen arbitrarily, being approximately equal to the ratio of the sampling frequency (500 Hz) over the high-frequency cut-off of the applied band-pass filter (55 Hz). We then averaged EEG amplitude across time points within each EEG segment, thus reducing the total number of time points by 10, as illustrated in Fig.1, based on one EEG time series as an example.

At the discretization stage, which was performed separately for each EEG channel, we divided the entire range of this EEG channel's amplitude into several bins, each assigned with a letter from the Latin alphabet. First, we defined two quantiles: Q1 and Q3, representing respectively the 25th and 75th percentiles in the range of EEG amplitude. Second, we calculated the Inter Quartile Range (IQR), which was the difference between Q3 and Q1. Finally, we defined the upper and lower boundaries, which were Q1 – 1.5 IQR and Q3 + 1.5 IQR, respectively. All values above the upper boundary and below the lower boundary were deemed outliers. We assigned two bins for the upper and lower outliers. The rest of the amplitude range was divided into 20 equally spaced bins. The entire range of EEG amplitude, separately for each channel, was thus discretized into 22 bins, denoted by the symbols from “a” to ”v”. Using this mapping, we assigned a symbol to each signal value obtained from PAA, as illustrated in Fig. 2. Each EEG time series was thus transformed into a string of symbols. e.i. "ababdc…". As a result of this procedure, all EEG signals were normalized and symbolized.

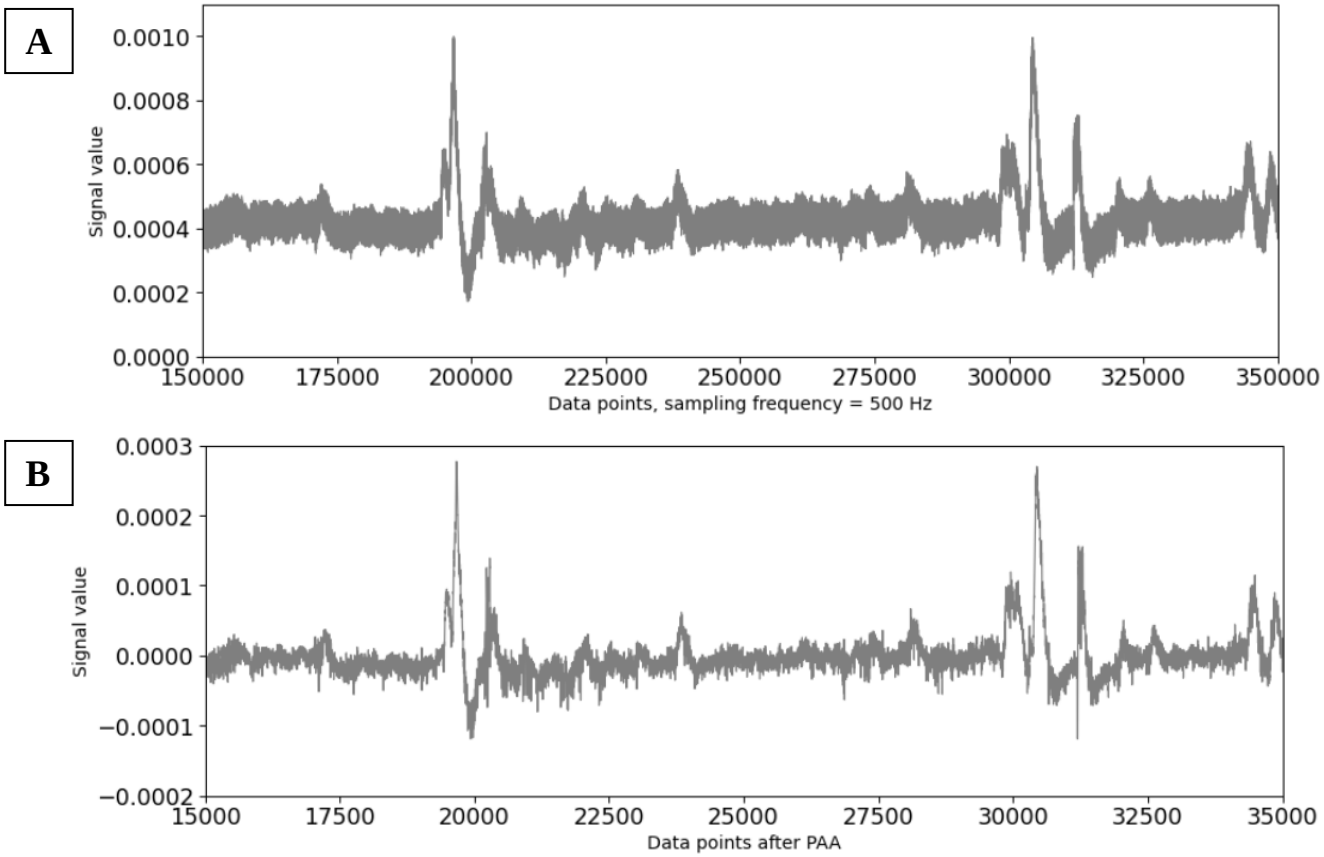


FIGURE 1: An example demonstrating one EEG signal (from the channel C3) before and after Piecewise Aggregate Approximation (PAA): (a) a 400-second fragment of the original EEG signal with a sampling frequency of 500 Hz before PAA, so its length is 200,000 data points; (b) the same fragment after applying PAA transformation with the window size of 10, so the new length is 20,000 data points.

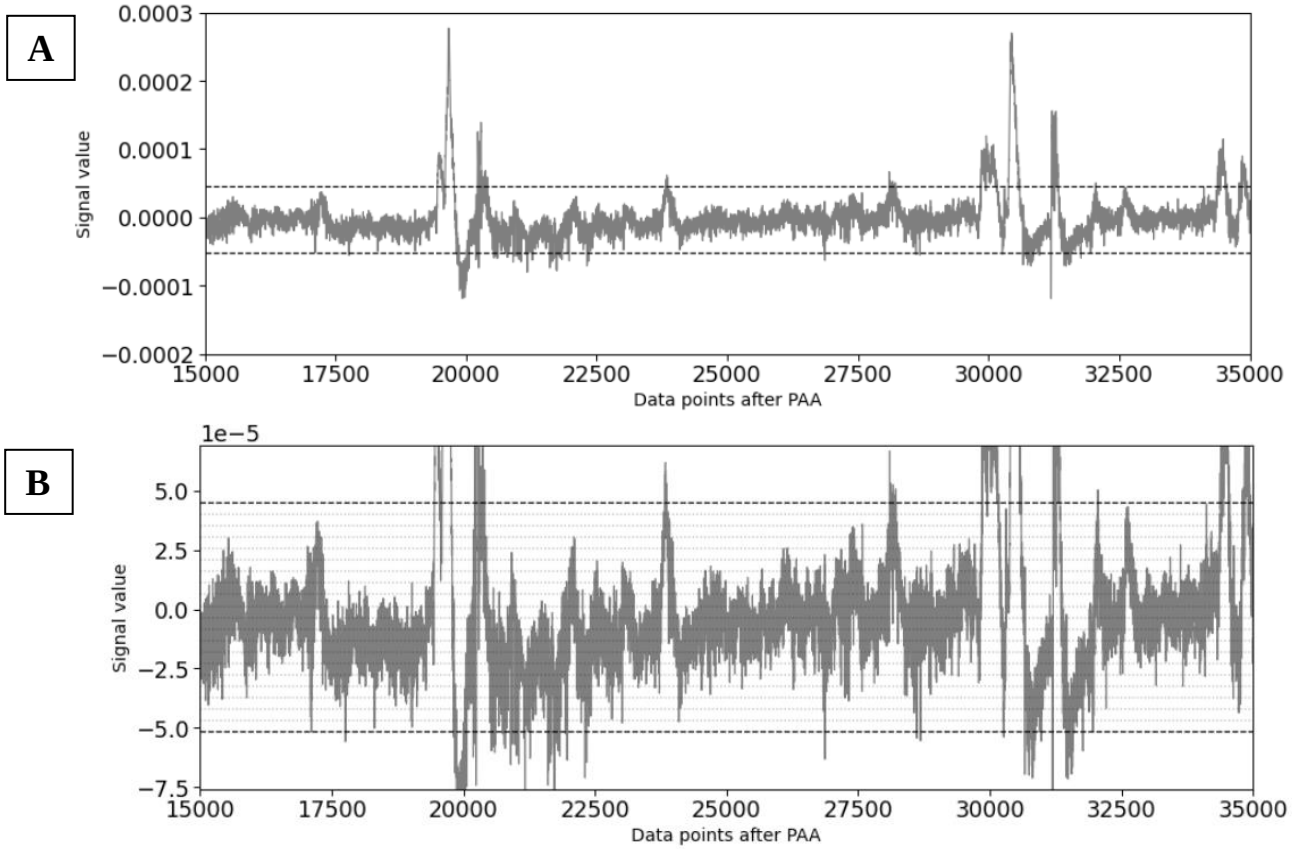


FIGURE 2: An example demonstrating the symbolization of EEG signals after the discretization stage, which is illustrated in Fig.1 : (a) the upper and lower boundaries (dashed lines) are defined as the Q1 - 1.5 IQR and Q3 + 1.5 IQR, which in turn defines two bins for the upper and lower outliers; (b) a 'zoomed-in' version (y-axis rescaled) of the same signal, as in Fig.2a, wherein we divide the amplitude range between the boundaries into 20 bins (dotted lines). Each bin, including those for outliers, is assigned a symbol from the Latin alphabet (from “a” to “v”, from bottom to the top).

### *D. Tokenization of EEG time series.*

Once EEGs were symbolized, we applied a Byte Pair Encoding (BPE) algorithm to split the strings of letters into tokens representing groups of letters standing together. BPE has been around for a long time (Gage, 1994); however, it has received much more recognition since it was applied in a study demonstrating its ability to handle rare and unknown word translation tasks (Sennrich et al., 2015). In general, the BPE algorithm addressed a problem of interpretation in the NLP domain. For example, compound words like "authorship" could be understood by NLP models as subwords "author-"and "-ship" because subwords "author" and "ship" alone often occur in human language. Extending the analogy, we aimed to find comparable patterns in symbolized EEG data. BPE takes all pairs of consecutive symbols, counts their occurrences in the text, and merges the most frequent pairs into new symbols. Then a newly merged symbol can be again merged with another symbol if the new pair occurs most frequently within a given iteration. In this way, the algorithm starts with a base vocabulary (symbols from “a” to “v”, in our case, that occur in the training dataset), learns rules for merging the basic symbols to form new symbols, and generates a new vocabulary of basic and merged symbols. We refer to these newly merged symbols as tokens. The algorithm (tokenizer) iterates until the vocabulary attains an a priori defined vocabulary size. Fig. 3 illustrates the basic steps performed by

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 <u>BA</u> B C D C <u>BA</u> B C D***

Step 2: most frequent is pair of BA and B, merge BA + B
***D C B <u>BAB</u> C D C <u>BAB</u> C D***

Step 3: most frequent is pair of BAB and C, merge BAB + C
***D C B <u>BABC</u> D C <u>BABC</u> D***

Step 4: most frequent is pair of D and C, merge D + C
***<u>DC</u> B <u>BABC</u> <u>DC</u> <u>BABC</u> D***

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

FIGURE 3. A schematic illustration of the byte-encoding algorithm (BPE). In Step 1, the algorithm finds the pair of symbols 'B' and 'A' to be the most frequent in the string, so BPE merges them into a new token, 'BA'. In the next step (Step 2), the token 'BA' makes the most frequent pair with the symbol 'B', so they are merged into a new token 'BAB'. In the end, we can describe our initial string with only four tokens: two 'BABC' tokens, two 'DC, one ‘B’, and one ‘D’, which compresses the original sentence. In our case, the input data is 5850 symbolized series, which makes it in total ~3,5 billion symbols (5850 EEGs, each having 20 channels, and 30000 symbols representing each channel).

the tokenizer.

To learn rules for merging symbols and find repeated tokens in the dataset, we trained the tokenizer on symbolic series of all EEG recordings in the entire dataset. We applied the open-source implementation of the BPE algorithm, as developed by the Hugging Face community (Hugging Face, BPE).

### *E. EEG features based on EEG tokens.*

Once the tokenizer had learned a vocabulary of tokens, we described each symbolic EEG series by a sequence of newly learned tokens, each representing a unique pattern of changes in EEG amplitude in the original EEG signal. We considered each token as a word composed of one or more letters and counted the number of tokens in each channel separately for each EEG segment, similar to the bag-of-words approach. The bag-of-words model is commonly used in NLP for text document classification, wherein each word's occurrence frequency is used as a feature for training a classifier (McTear et al., 2016). We weighted the tokens' occurrence frequency by their length and normalized it with respect to the total length:

$$\text{Token's occurrence frequency, \%} = \frac{\text{Number of token appearances} \times \text{Token length}}{\text{Total length of the series}} \times 100$$

Note that for a given channel, these values sum up to exactly 100%. As we applied the tokenizer separately to each EEG channel, the tokens' occurrence frequency was calculated across the pool of tokens generated for a given channel, as illustrated in Fig. 4. Thus, the total number of tokens across all EEG channels was approximately equal to the vocabulary size of the tokenizer multiplied by the number of channels. We say approximately as there was no guarantee that the same tokens would appear in all channels. So, some channels may lack tokens present in the vocabulary. As a result, the EEG dataset had a size of 5850 samples with about 32000 features, representing the tokens' occurrence frequencies across 20 channels.

We also considered situations wherein tokens had the same shape but represented different EEG amplitudes, as exemplified in Fig. 5. Specifically, two tokens, 'bdc' and 'dfe' in Fig. 5a have the same shape, but they are shifted across the amplitude axis. Formally, these two patterns of changes in EEG amplitude are represented by different combinations of letters. We converted all the tokens into their "relative" form (Fig. 5b) by calculating the distance between the adjacent symbols in terms of the number of bins between them. For example, in the token 'bdc', the distance between the letters 'b' and 'd' is two bins up, and that between 'd' and 'c' is one bin down. Thus, both tokens in Fig. 5a had the same form [+2, -1] (Fig. 5b). In our study, we used the terms of "symbolic

| | 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 4: A schematic illustration of the feature space to describe tokenized EEG signals. Each row is associated with one EEG segment. Each column represents a feature associated with one token. Its value is the token's occurrence frequency for a given EEG channel. The tokens are channel-specific, so columns are ultimately named in the format 'channel_token': for example, 'C4_BABC' or 'O2_BABC'. Note that the tokens' occurrence frequencies sum up to 100 % for each EEG channel.

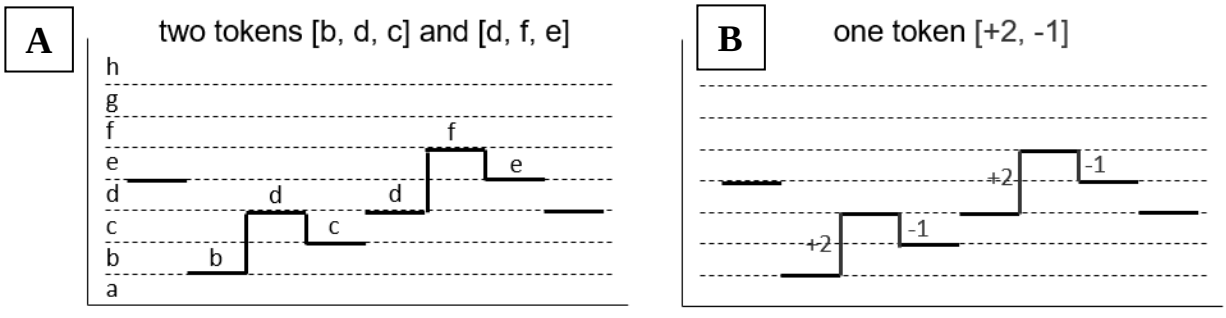


FIGURE 5: Example of transition from two symbolic tokens to one relative form token. (a) The first subplot shows the example of two tokens, 'BDC' and 'DFE', which have the same shape but are assigned different symbols because of different signal values. (b) The second subplot highlights the distance between symbols, counted in the number of bins, like the distance between B and D is two bins, so the resulting form of both symbolic tokens becomes [+2, -1].

tokens" and "relative tokens" to designate the original tokens and tokens representing relative changes in EEG amplitude, respectively.

The transition from the absolute symbolic form to the relative form decreased the number of features in our dataset from approximately 32,000 to 8,000. To compute the occurrence frequencies of tokens representing relative changes in the EEG amplitude, we summed up the tokens' occurrence frequencies across those tokens that had the same relative shape. This was done separately for each EEG channel. For example, if token 'bdc' in channel O2 occurred with a frequency of 0.015, and the token 'dfe' in the same channel O2 occurred with a frequency of 0.01, the corresponding relative token with the shape O2 [+5, -2] would have the occurrence frequency of 0.025. Note that as before, all frequencies within a given channel summed up to 100%.

### *F. Tokens-based EEG features and their relationships with age.*

We correlated the tokens' occurrence frequencies with patients' biological age to validate our pipeline based on the symbolization of EEG signals, and demonstrated a physiologically-relevant representation of tokens. We applied two approaches. First, we trained a classical (non-neural network) machine learning model to predict patients' age using tokens' occurrence frequencies as new EEG features to assess these features' predictive power as a whole. Second, we correlated each EEG feature of interest with age univariately to find features most sensitive to changes in age.

**Correlations between tokens' occurrence frequencies and age.** We explored correlations between each token's occurrence frequency and age across subjects. Distance correlation (Richards, 2017) was designed to reflect functional, nonlinear associations between two variables. Separately for each token and EEG channel, we calculated the distance correlation coefficient between this token's occurrence frequency and patients' age. We also analyzed how this measure varied across EEG channels. For a few tokens with the strongest correlation, we explored how their occurrence frequency changed with age and what patterns in EEG signals they represented.

**Age prediction with a machine learning model.** We tested the capacity of the new EEG features to predict the patient's age. Specifically, we tested a random forest regression model, using the tokens as features, the tokens' occurrence frequencies as the predictor variables, and the patients' age as the target variable.

The dataset was split into train and test subsets (80% of EEG segments were used for training and 20% for testing). We applied the cuML RAPIDS implementation (RandomForest Regressor, RAPIDS) of the Random Forest model to utilize high-performing training on GPU (Tesla T4 16GB). The model had the following parameters: the number of decision trees in the forest was equal to 1,000; the tree depth or the number of nodes from the root of a tree to the final leaf was set to 16 . Other parameters were kept default . We evaluated the performance of the age prediction model with three metrics: (a) Mean Absolute Error (MAE) between the actual and predicted age, (b) Pearson correlation coefficient between the predicted and actual age, and (c) Explained variance score, which measures the proportion to which a model accounts for the variation of a dataset. The model was trained and evaluated twice for two sets of features: symbolic and relative tokens.

## III. RESULTS

Under the framework of random forest regression for age prediction, the symbolic tokens extracted from raw EEG signals provided reasonably good performance. The model has predicted patients' age with an MAE = 15.6 years (Fig. 6). The Pearson correlation coefficient between the predicted and actual age in the test sample was 0.54, whereas the explained variance score was 0.3.

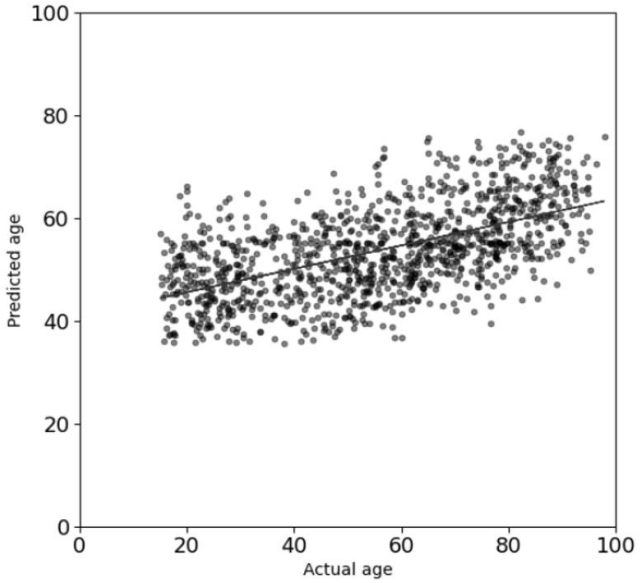


FIGURE 6. The scatterplot shows relationships between the actual and predicted age on the dataset with relative tokens. Each dot represents an EEG sample from the test subset with a fitted linear regression line. The line should have a 45-degree slope for ideal model predictions with MAE = 0 years and Pearson correlation = 1. Our regression line is far from having a 45-degree slope, which is a typical statistical effect called "regression to the mean".

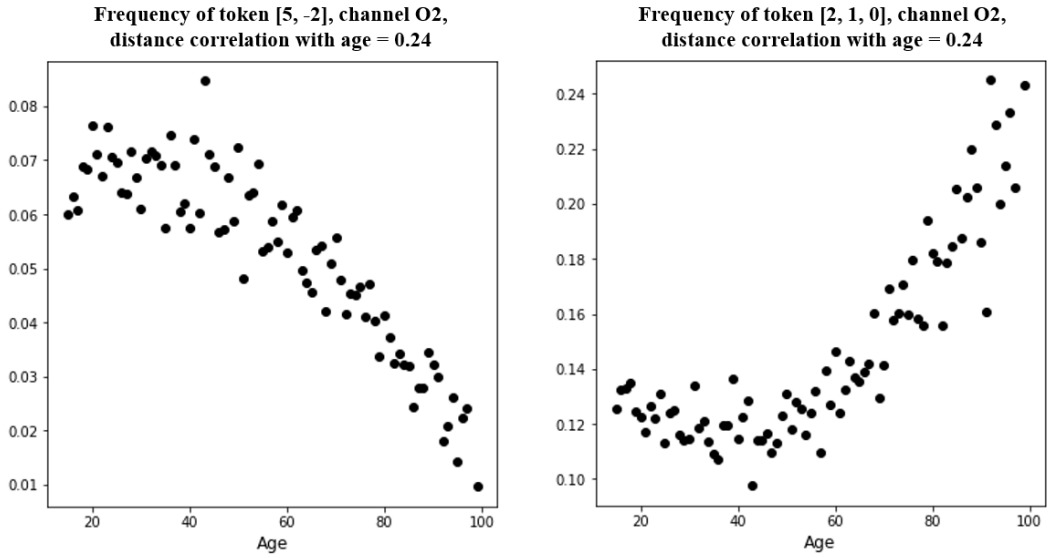


FIGURE 7: Relationships between relative token occurrence frequency and age. Two relative tokens were identified as having the highest distance correlation values: [+5, -2] (left) and [+2, +1, 0] (right), both representing changes in EEG amplitude on the channel O2. The entire age range was divided into non-overlapping age groups, with a moving step of one year. Each dot on these scatterplots represents the mean occurrence frequency averaged across subjects within each age group.

Compared to the symbolic tokens, the random forest regression model applied to the relative tokens demonstrated qualitatively very similar performance in terms of MAE, correlation, and explained variance. We note, however, that a smaller number of relative tokens compared to original symbolic tokens (approximately 8,000 vs. 32,000) significantly improved the computational time required for the model training: about 40 minutes versus 2 hours.

We employed the distance correlation metric to find out the most influential tokens and explored how their occurrence frequency changes with age. We calculated the distance correlation between tokens' occurrence frequency and patients' age separately for each token. The tokens with the highest correlations (with a threshold at the 99.99-percentile) had their distance correlation values in a range between 0.25 and 0.29 for the symbolic tokens and 0.23 to 0.24 for relative tokens. We identified the two most influential relative tokens and explored how their occurrence changed with age. Then, we grouped all the patients from the training dataset into non-overlapping one-year-long age categories. Fig. 7 shows the functional relationships between the tokens' occurrence frequency and age, wherein each dot represents the tokens' mean occurrence frequency, which was averaged across patients within each age category. As can be seen from Fig.7, these relationships are nonlinear in general, with positive or negative trends depending on the age range.

We have also checked how the distance correlation between tokens' occurrence frequency and age varied across the spatial organization of EEG channels. On average, some EEG channels expressed stronger correlations. We computed the median and mean values of the correlation coefficients across all tokens within each EEG channel. We found that tokens extracted from EEG recorded in the frontal and occipital areas have a higher median distance correlation. Visually, the distributions of median or mean correlation values across channels tended to preserve a spatial symmetry between the left and right hemispheres (Fig. 8). Note that this symmetry was not modeled by the proposed workflow of analyses.

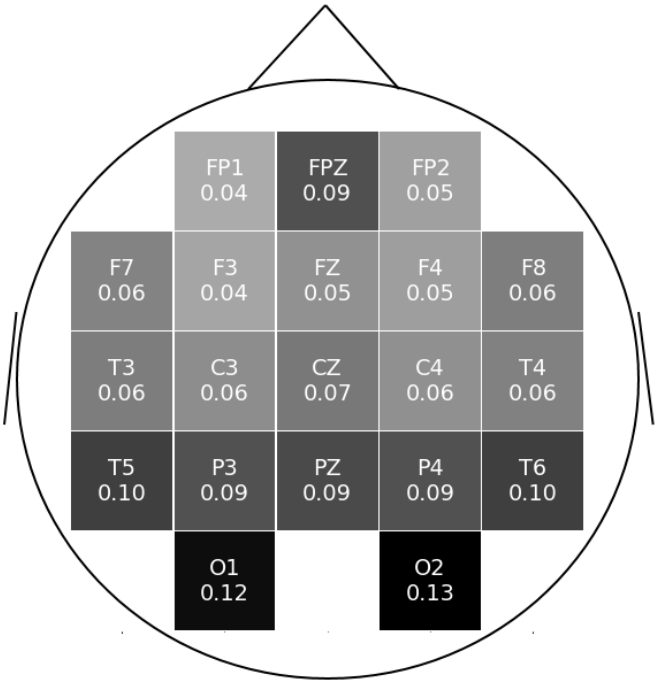


FIGURE 8. Distribution of the median distance correlations between tokens' occurrence frequency and patient's age. The median value was computed separately for each EEG channel across all tokens. Each cell shows an EEG channel name and its median correlation value. The darker areas stand for higher correlations.

From the EEG symbolization stage, we know the exact location tokens and can trace them all in the original EEG signal. Our method thus allows one to find the EEG fragments corresponding to the tokens that correlate the most with age (Fig. 9). Notably, these EEG segments are relatively short. For example, with the original sampling frequency of 500 Hz and the window size of 10 data points used for defining a symbol, the token of three symbols has a duration of 3 x 20 = 60 milliseconds.

## IV. Discussion

In our study, we applied an NLP-based pipeline for extracting features from clinical EEG recordings to their subsequent classification. We tested this approach using an extensive sample of routine clinical EEG scans recorded in a large pool of patients within a very wide age range. First, with minimal preprocessing, we converted EEG signals into symbolic series. The byte-pair encoding (BPE) algorithm then extracted the most frequent patterns (tokens) from newly reconstructed symbolic series. To validate our approach, we used the extracted tokens as new EEG features and applied a random forest regression model to predict patients' age. We also examined functional relationships between tokens occurrence frequency and patients' age and found that these relationships may follow a U-shape pattern, depending on the age. Spatially, the most strong correlations were expressed by EEG channels from the prefrontal and occipital areas.

We tested a minimal preprocessing of EEG signals recorded in a complex clinical environment, using a classical (non-neural) model without fine-tuning the model parameters. Still, we demonstrated the presence of relatively high correlations between tokens and age and obtained results that were comparable with other studies. Our mean absolute error (MAE) in correlating the actual and predicted age was relatively low compared to other studies on age prediction from EEG. For example, two studies reported MAE of 7.6 and 6.9 (Sun et al., 2018, Zoubi et al., 2018). The highly heterogeneous nature and size of our EEG sample may explain the lower performance of our approach compared to other studies. Our workflow of analyses processed EEG as it was recorded in real-world clinical practice scenarios. More specifically, our analysis was based on 10-minute long EEG segments, whereas comparable approaches utilized 30-60 seconds epochs (Sun et al., 2018, Zoubi et al., 2018). We analyzed virtually all EEG scans recorded and evaluated in the process of the diagnostic workup in one hospital without any selection bias. These EEG recordings were clinically classified as either normal or abnormal. Our population included both in-patients (they are required to stay in the hospital overnight) and out-patients (no such requirement). This is a highly heterogeneous population, with various diagnoses and a full spectrum of comorbidity levels. Typically, these patients take various medications, which are known to impact EEG (Vakorin et al., 2021). Also, we did not use hyperparameter tuning techniques for our models, as our primary goal was to demonstrate the feasibility of an NLP-based approach for classifying clinical EEG.

The proposed workflow of analyses included two main stages: symbolization of EEG records and their subsequent tokenization. Symbolization or transformation of EEG time series into symbol (letters) sequences is not new (Liu et al., 2006). A number of studies applied this approach to

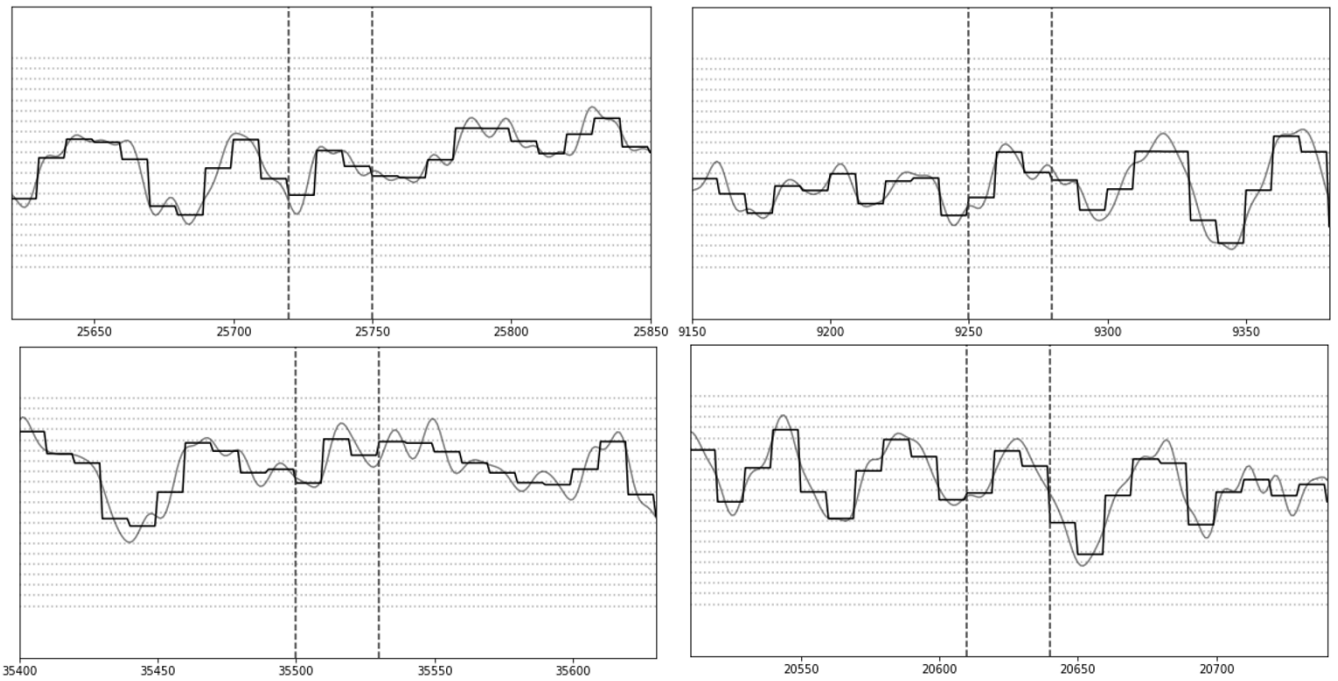


FIGURE 9: Four different EEG segments from original EEG signal with two annotations, indicating the beginning and end of the token [+5, -2] in channel O2. The original EEG signals are shown in a light-gray, whereas and PAA-transformed signals are shown in black. Dotted lines represent the margins of the EEG amplitude bins, which correspond to different symbols (letters). On each subplot, the letters in all tokens first go five bins up, and then two bins down, although at different amplitude levels (along the y-axis). Each subplot has a duration of 460 milliseconds, with the marked token's duration of 60 milliseconds.

characterizing the signal complexity of EEG dynamics with metrics such as symbolic entropy (Hussain et al., 2017), (Liu et al., 2006), in various applications, e.g., anesthetics effects on cortical information flows (Jordan et al., 2011). Typically, the symbolization stage in the analysis performed by those studies was used to define a single value, which characterizes the entire time series, such as the entropy of a given signal. At a later stage, EEG classification or group analysis was performed based on the macro characteristics of EEG signals, which by design prevents tracing group differences back to original EEG dynamics.

In contrast to symbolization, tokenizing the symbols or grouping them into "words" is novel in the EEG literature, and this contributes significantly to solving one of the key issues in machine learning applications in healthcare, namely, interpretability (Shmueli, 2010). The proposed method of feature extraction with BPE ensures that the extracted features, namely tokens or ultimately short patterns in EEG amplitude, are interpretable. Here we define interpretability as the pipeline's ability to directly point to the most sensitive EEG waveforms and their exact locations in the original signal. A combination of symbolization and tokenization of EEG time series allows one to focus on relatively short patterns of changes in EEG amplitude, which can be of high relevance to a target variable of interest. Ultimately, this functionality would allow physicians to focus on the EEG patterns in the original recording, which corresponds to a clinically-relevant parameter. This can be particularly important in the context of inter-rater variability reported in evaluating clinical EEG. Sharp transients in EEG, including interictal epileptiform discharges, vary significantly in their morphological properties, such as voltages, durations, slopes, areas, and across-channel correlation, which may not be interpreted reliably by different physicians (ref jing). One of the advantages provided by BPE tokenization is that EEG transients can be potentially identified automatically in an unsupervised manner. Importantly, our method can directly identify very short clinically relevant EEG waveforms and characterize their morphological features, such as duration, voltage amplitude, and changes in the slope, without significant modifications.

In general, the proposed approach offers several advantages for EEG classification. Potentially, it can handle EEG recordings of arbitrary duration, as it was originally designed to classify unstructured texts of varying length. The method does not require advanced EEG preprocessing and finds local patterns (in the temporal domain) in raw EEG in an unsupervised manner. The feature extraction method can be applied to other classification tasks besides age prediction. Importantly, the proposed approach can identify short clinically-relevant events, such as interictal epileptiform discharges. The local traceable features can be potentially utilized in decision support systems to highlight clinically-relevant EEG segments for further examination by neurologists with conventional methods.

## REFERENCES


[1] ACNS, "American Clinical Neurophysiology Society: Guidelines and Consensus Statements.," https://www.acns.org/practice/guidelines, 2022.

[2] G. Schneider et al., "Quality of perioperative AEP—variability of expert ratings," British Journal of Anaesthesia, vol. 91, no. 6, pp. 905–908, Dec. 2003, doi: 10.1093/bja/aeg280.

[3] A. A. Asadi-Pooya and M. R. Sperling, "Normal Awake, Drowsy, and Sleep EEG Patterns That Might Be Overinterpreted as Abnormal," Journal of Clinical Neurophysiology, vol. 36, no. 4. 2019. doi: 10.1097/WNP.0000000000000585.

[4] J. Y. Kang and G. L. Krauss, "Normal Variants Are Commonly Overread as Interictal Epileptiform Abnormalities," Journal of Clinical Neurophysiology, vol. 36, no. 4, pp. 257–263, Jul. 2019, doi: 10.1097/WNP.0000000000000613.

[5] S. v. Mathias and M. Bensalem-Owen, "Artifacts That Can Be Misinterpreted as Interictal Discharges," Journal of Clinical Neurophysiology, vol. 36, no. 4. 2019. doi: 10.1097/WNP.0000000000000605.

[6] M. Teplan, "FUNDAMENTALS OF EEG MEASUREMENT M . Teplan," Measurement Science Review, vol. 2, no. Section 2, 2002.

[7] G. Pfurtscheller and C. Neuper, "Motor imagery and direct brain-computer communication," Proceedings of the IEEE, vol. 89, no. 7, 2001, doi: 10.1109/5.939829.

[8] R. G. Andrzejak, K. Lehnertz, F. Mormann, C. Rieke, P. David, and C. E. Elger, "Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state," Physical Review E, vol. 64, no. 6, p. 061907, Nov. 2001, doi: 10.1103/PhysRevE.64.061907.

[9] V. A. Vakorin, S. M. Doesburg, L. da Costa, R. Jetly, E. W. Pang, and M. J. Taylor, "Detecting Mild Traumatic Brain Injury Using Resting State Magnetoencephalographic Connectivity," PLOS Computational Biology, vol. 12, no. 12, p. e1004914, Dec. 2016, doi: 10.1371/journal.pcbi.1004914.

[10] D. Kim and K. Kim, "Detection of Early Stage Alzheimer's Disease using EEG Relative Power with Deep Neural Network," in Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, 2018, vol. 2018-July. doi: 10.1109/EMBC.2018.8512231.

[11] U. R. Acharya, S. L. Oh, Y. Hagiwara, J. H. Tan, H. Adeli, and D. P. Subha, "Automated EEG-based screening of depression using deep convolutional neural network," Computer Methods and Programs in Biomedicine, vol. 161, 2018, doi: 10.1016/j.cmpb.2018.04.012.

[12] Y. Guo, K. Friston, A. Faisal, S. Hill, and H. Peng, "Brain informatics and health: 8th international conference, BIH 2015 London, UK, august 30 – september 2, 2015 proceedings," in Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 2015, vol. 9250.

[13] K. G. van Leeuwen, H. Sun, M. Tabaeizadeh, A. F. Struck, M. J. A. M. van Putten, and M. B. Westover, "Detecting abnormal electroencephalograms using deep convolutional networks," Clinical Neurophysiology, vol. 130, no. 1, 2019, doi: 10.1016/j.clinph.2018.10.012.

[14] A. Craik, Y. He, and J. L. Contreras-Vidal, "Deep learning for electroencephalogram (EEG) classification tasks: A review," Journal of Neural Engineering, vol. 16, no. 3. 2019. doi: 10.1088/1741-2552/ab0ab5.

[15] I. Stancin, M. Cifrek, and A. Jovic, "A Review of EEG Signal Features and Their Application in Driver Drowsiness Detection Systems," Sensors, vol. 21, no. 11, p. 3786, May 2021, doi: 10.3390/s21113786.

[16] R. T. Schirrmeister et al., "Deep learning with convolutional neural networks for EEG decoding and visualization," Human Brain Mapping, vol. 38, no. 11, 2017, doi: 10.1002/hbm.23730.

[17] T. B. Brown et al., "Language models are few-shot learners," in Advances in Neural Information Processing Systems, 2020, vol. 2020-December.

[18] M. Wang, D. Cao, L. Li, S. Li, and R. Ji, "Microblog Sentiment Analysis Based on Cross-media Bag-of-words Model," in Proceedings of International Conference on Internet Multimedia Computing and Service - ICIMCS '14, 2014, pp. 76–80. doi: 10.1145/2632856.2632912.

[19] M. Horak, S. Chandrasekaran, and G. Tobar, "NLP Based Anomaly Detection for Categorical Time Series," Apr. 2022.

[20] N. Tavabi and K. Lerman, "Pattern Discovery in Time Series with Byte Pair Encoding," May 2021.

[21] L. Hussain, S. A. Shah, W. Aziz, S. N. H. Bukhari, K. J. Lone, and Q.-A. Chaudhary, "Analyzing the dynamics of sleep electroencephalographic (EEG) signals with different pathologies using threshold-dependent symbolic entropy," Waves in Random and Complex Media, vol. 31, no. 6, pp. 2337–2354, Nov. 2021, doi: 10.1080/17455030.2020.1743378.

[22] D. Jordan, S. Paprotny, E. Kochs, G. Schneider, and Research Group on Brain Mechanisms of Consciousness and Anaesthesia, "Symbolic transfer entropy indicates changes of cortical flow of information between consciousness and propofol-induced unconsciousness: 7AP1-4," European Journal of Anaesthesiology, vol. 28, p. 97, 2011.

[23] Ying Liu, Lisha Sun, Yisheng Zhu, and P. Beadle, "Novel Method for Measuring the Complexity of Schizophrenic EEG Based on Symbolic Entropy Analysis," in 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference, 2005, pp. 37–40. doi: 10.1109/IEMBS.2005.1616336.

[24] H. Nahrstaedt, "PyEDFlib, EDF/BDF Toolbox in Python," https://pyedflib.readthedocs.io/, 2022.

[25] E. Keogh, K. Chakrabarti, M. Pazzani, and S. Mehrotra, "Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases," Knowledge and Information Systems, vol. 3, no. 3, 2001, doi: 10.1007/pl00011669.

[26] P. Gage, "A New Algorithm for Data Compression," The C Users Journal, 1994, doi: 10.5555/177910.177914.

[27] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016 - Long Papers, 2016, vol. 3. doi: 10.18653/v1/p16-1162.

[28] HuggingFace, "Byte-Pair Encoding tokenization," https://huggingface.co/course/chapter6/6, 2021.

[29] M. McTear, Z. Callejas, and D. Griol, The Conversational Interface. Cham: Springer International Publishing, 2016. doi: 10.1007/978-3-319-32967-3.

[30] D. St. P. Richards, "Distance Correlation: A New Tool for Detecting Association and Measuring Correlation Between Data Sets," Aug. 2017.

[31] nvidia, "Random Forest Regressor, RAPIDS," https://docs.rapids.ai/api/cuml/stable/api.html?highlight=random%20forest%20regressor#cuml.ensemble.RandomForestRegressor, 2020.

[32] O. al Zoubi et al., "Predicting age from brain EEG signals-a machine learning approach," Frontiers in Aging Neuroscience, vol. 10, no. JUL, 2018, doi: 10.3389/fnagi.2018.00184.

[33] V. A. Vakorin et al., "Alterations in coordinated EEG activity precede the development of seizures in comatose children," Clinical Neurophysiology, vol. 132, no. 7, pp. 1505–1514, Jul. 2021, doi: 10.1016/j.clinph.2021.03.015.

[34] L. Hussain et al., "Symbolic time series analysis of electroencephalographic (EEG) epileptic seizure and brain dynamics with eye-open and eye-closed subjects during resting states," Journal of Physiological Anthropology, vol. 36, no. 1, 2017, doi: 10.1186/s40101-017-0136-8.

[35] G. Shmueli, "To Explain or to Predict?," Statistical Science, vol. 25, no. 3, Aug. 2010, doi: 10.1214/10-STS330.