Dynamically Decoding Human Physiological Behaviors from Intracranial Field Potentials

A DISSERTATION PRESENTED
BY
MANANA HAKOBYAN
TO
THE DEPARTMENT OF INSTITUTE FOR APPLIED COMPUTATIONAL SCIENCE

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
MASTERS'S
IN THE SUBJECT OF
DATA SCIENCE

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
MAY 2022

Thesis advisor: Professor Gabriel Kreiman Manana Hakobyan

Dynamically Decoding Human Physiological Behaviors from Intracranial Field Potentials

Abstract

The attempt to decode the human brain using computers is not novel, however, doing it dynamically in an uncontrolled environment with many external confounding factors has been deemed to be very challenging computationally, and hence, yet to be explored in depth. This study aims to predict the continuous human physiological behaviors using machine learning and invasively recorded intracranial field potentials received through electroctrocortigography (ECoG) procedure from the brain surface in an uncontrolled real life setting. After a rigorous feature engineering process I showcase that the well-defined behaviors such as sleeping, eating and video gaming can be decoded with greater than 0.95 AUCs, and the noisier behaviors such as movements, spoken and heard speech are decoded with AUCs higher than 0.80. To ensure that the classification results are reliable I run a series of experiments with different controls and find that despite the drop in AUCs the behaviors are still robustly classified better than the random for all of the tests. I also explore the specific brain regions responsible for the high performance of the classifications. Not only does this research show that it is possible to classify twelve natural continuous human behaviors with high performance, it also confirms many of the prior literature findings which state that certain brain region activities correspond to specific human physiological actions.

Contents

Listing of figures

To my sister.

Acknowledgments

I AM IMMENSELY GRATEFUL to Professor Gabriel Kreiman who allowed me to work in his lab and advised me throughout this challenging process. My sincere gratitude goes to Professor Pavlos Protopapas for being my reliable support system in IACS. And of course, I want to acknowledge all the people who have been by my side during the two years at Harvard including my parents, my brother, Armen and my incredible friends.

0
Introduction

Decoding brain activity in relation to physiological behaviors can be the key to advancing medical practices to solve many health-related problems which yet to be addressed. For example, understanding how and which brain parts correspond to certain eating behaviors can lead to successful medical therapies for curing eating disorders. Additionally, the ability to predict how and when the person wants to move and which type of movement they want to undertake from the neuronal activity can be translated into smart wearable devices - brain computer interfaces (BCI) - for giving mobility to those who lost their limbs. However, the successful application of any classification model in brain computer interfaces requires that the underlying research, and hence, the technology is robust to the challenges and the noise of the dynamic world. My research complies with the modern BCI demands of studying the brain in a dynamic uncontrolled environment making its results directly applicable to real-life problems.

In this study I work with data received from three human subjects who underwent an electrocorticography procedure prior to surgery for severe epilepsy and had electrodes implanted on the surface of their brain for at least a day. During this time the patients have also been video and audio-recorded making it possible to know what they have been doing at any second. The knowledge of their physiological behaviors during the time when their brain activity has been recorded makes it possible to formulate the challenge of this study as a supervised learning problem which is the framework used throughout the paper. To set the grounds for a successful demonstration of the series of results, the rest of the introduction is dedicated to defining the common concepts as they relate to the neuroscience context of this problem and shedding light on the previous literature for this study specifically.

0.1 Neuroscience Context and Brain Computer Interface

Neuroscience, the scientific field which is concerned with understanding the brain's physical, chemical and electrical structure, has been reaching major milestones in the past decades due to the phenomenal technological advances.[12] In the 1970s the research on Brain Computer Interfaces (BCI) started in the University of California leading to a new wave of innovations in the field.[27] BCI is a computer interface that establishes a full pipeline of acquiring the brain signals, analyzing and translating them into commands that can be carried out by an output device. In 2004, Wolpaw et al showed that using computers it is possible to translate the electrical brain waves into physical activity

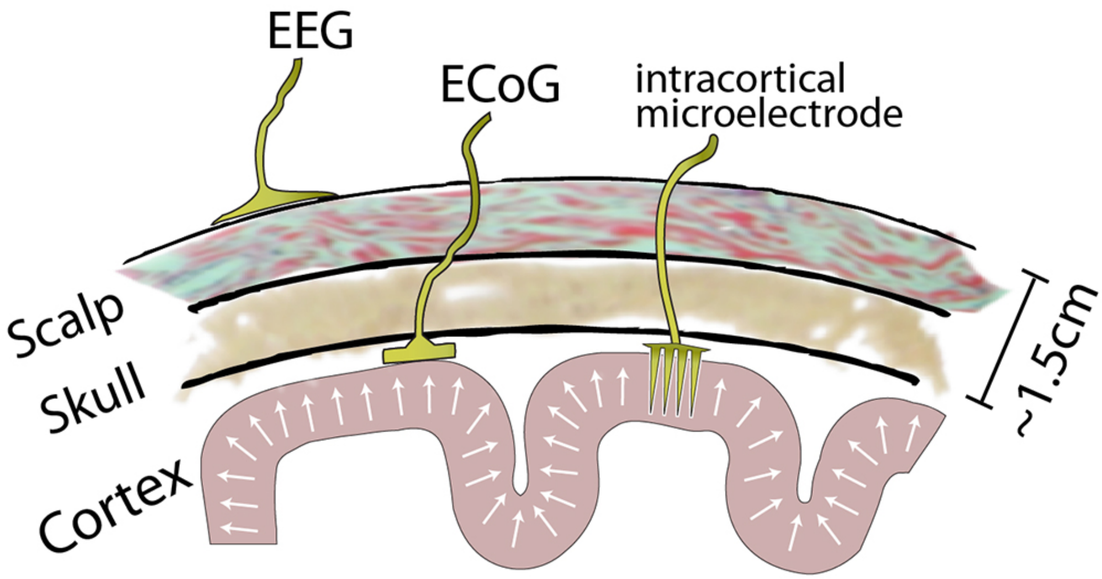

Figure 1: The figure demonstrates the different methods of recording field potentials from the brain and their exact placements.

such as moving a cursor on the screen.[45] A study published in August, 2021 demonstrated the use of nearly 50 autonomous wireless microimplants for recording rodent's neural activity which was a step towards addressing the brain and spinal injuries from within.[24] The most recent survey on data science and machine learning techniques in BCIs stresses the importance of the computational tools in handling brain activity data, feature extraction and classification, as the new problems posed for BCIs require working with big volumes of data and implementing sophisticated computational algorithms.[35] The current aim of BCIs is to restore the physical functionalities which are caused by neuromuscular disorders, but it also largely implies understanding the biology of the brain.

The improvements of non-invasive and invasive data collection methods driven by BCIs ended the cycle of seeing the brain as a blackbox mystery caused by the limitations of objectively studying its activity and biological structures. One such way to look into the brain is through Electrocorticography (ECoG), also known as intracranial electroencephalography (iEEG), which is defined to be the process of recording the electrical activity invasively by placing the electrodes in direct con-

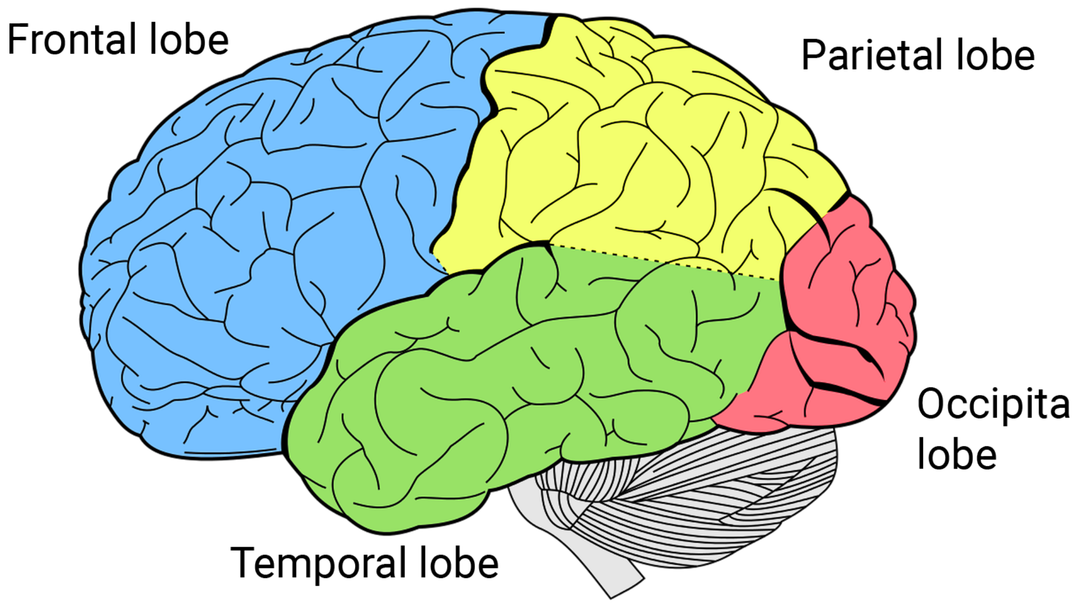

Figure 2: Shown above are the four lobes in human brain and their relative placements.

tact with the surface of the brain.[40 Norcliffe-Kaufmann et al.] Not only do the intracranial field potentials (IFPs) received through ECoG have excellent spatial and spectral resolution, but also the ability to maintain high signal-to-noise ratio due to the close proximity of electrode placements to the source of activity.[26] Since the process of obtaining iEEG data requires open-cut placement of implants in the scalp, currently, the procedure is mainly used in the clinical setting for pre-surgical epileptic patients.[6] The IFPs are analyzed during the pre-surgical procedure to locate the epileptic foci and after the surgery to assess the completeness of the treatment. Other major ways of accessing the brain activity include but are not limited to Electroencephalogram (EEG) which requires placing electrodes non-invasively on the scalp; and through intracortical neural microelectrodes which are implanted directly in the cortex for both recording and stimulating neural activity.

The human brain has two hemispheres and 4 sections associated with them which are called lobes: the frontal lobe which is generally associated with personality, decision making and movement; the parietal lobe which supports humans to identify objects, understand spatial relationship, respond to touch and interpret spoken language; the occipital lobe which is linked to vision; and the temporal lobe which is responsible for short-term memories, speech, rhythm and smell.[2] The outermost neural tissue sheet of our brain is called the cerebral cortex and it is also exactly where the electrodes for ECoG procedure are implanted.[3] The cerebral cortex has 10^{10} neurons and is deemed to be the place where higher-order computations are performed.[34] A cerebral cortex has a wrinkled appearance and consists of gyri(the ridges or bumps) and sulci(the grooves).[1]

0.2 Literature Review

The net voltages recorded by ECoG and the various frequencies of their oscillations have been linked to distinct human physiological behaviors, serving as good grounds for researching how to decode those behaviors from the brain activity further.[10] Using ECoG data with paired video recordings obtained from implanted electrodes in pre-neurosurgical diagnostics of epilepsy patients Deris et al. started linking the spontaneous real-world speech to so-called cognitive "idea" units (IUs).[15] Deris et al. used invasive neuronal recordings for brain activity and video-recordings for tracking the dynamic human behavior, but the main aim of their study remained to be very different from mine, as their goal was to explore human speech. ECoG data are also used widely for studying the brain parts and their functionalities, for example the experiment by Tang et al. where 15 epileptic patients were studied while they were performing a Stroop task revealed that the four regions of the frontal cortex responded to altered colors and word meanings in their task.[39]

In 2011 Z. Wang et al published results about decoding the onset and direction of movements using ECoG, showing that the ECoG signals contain valuable information for understanding the initiations of different movement types.[44] iEEG data were shown to be valuable also for understanding mental states and predicting mental fatigue by Yao et al in 2020.[46] The particular study with Yao et al worked with non-human primates(NHP) where they tracked the NHPs over extended periods of time and had them execute mentally demanding tasks in a controlled environment. While the study established good grounds for the possibility of exploring mental fatigue from within, it struggled to generalize to humans without further research and investigations with human subjects.

N. X. R. Wang's research on clustering the ECoG data via hierarchical clustering and then using computer vision with video-recordings for characterizing the general clusters with human behavior related features such as speech and movement is the closest literature to my study.[43] However, even though the studies have a similar data set up, uncontrolled experimental design and similar motivations of decoding human behaviors from ECoG, their adopted methodology is different from the one in this study. In fact, the authors state themselves that their approach of implementing unsupervised learning and computer vision for extracting features from the video recordings lack the specificity and precision in decoding the behaviors, which I hope to add in my study with a supervised learning problem-solving framework.

Finally, according to a 2021 survey of neural decoding papers the state of machine learning usage in BCI related research using EEG data follows the trend of engineering domain-specific informative features and keeping the algorithms simple as long as they achieve good classification results.[36] In fact, Random Forests, k-NNs, SVMs and simple CNNs are the best performing models for most of the decoding tasks. In this study I also adopt similar approach of performing a rigorous feature engineering and using more interpetable models.

A typical neuron makes about ten thousand connections to neighboring neurons. Given the billions of neurons, this means there are as many connections in a single cubic centimeter of brain tissue as there are stars in the Milky Way galaxy.

David Eagleman

1

Data and Methods

1.1 Data

The data used in this paper are recorded from three human subjects at Boston Children's Hospital (BCH) and were approved for research by BCH Institutional Review Board. All of the subjects have been patients in the hospital undergoing a neurosurgical procedure to remove the seizure foci in the brain. The patients vary in age and gender including one female and two males and ranging from age 11 to 32. In order to intraoperatively confirm the location and extent of epileptic tissue the patients were implanted with electrodes in direct contact with the surface of the cerebral cortex - a procedure known as ECoG.[37] The implanted electrodes had a diameter size of 1.3 mm. Hence, the data of this research are the electrical activity, or more formally the intracranial field potentials (IFP), received from the invasive electrodes measured in microvolts (μV). The sampling rate of the original data collection was 500 Hz, but later the data were undersampled to roughly 250 samples per second to make it easy and efficient when it comes to storage and distribution. The data were also notch filtered at 20 Hz as well as 60 Hz and harmonics as shown in the Figure 1.1. This was done to get rid of any additional interference in the signal recording process such as the AC current in the electrical devices.[14]

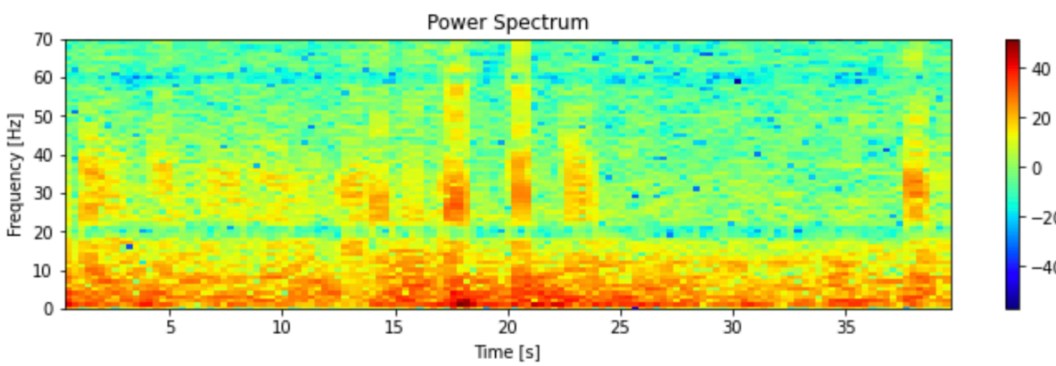

Figure 1.1: The figure shows the power spectrum of the field potentials for Subject 2.

In modern EEG/iEEG, there are two types of commonly used electrode configurations: unipolar and bipolar. In a bipolar configuration, two electrodes are placed on the head, and the resulting signal is the difference between those two electrodes.[25] In a unipolar configuration, multiple electrodes are placed at different locations and one of them serves as a reference for the rest of the electrodes. The majority of the implanted electrodes for all of our patients have a reference electrode which could be used for bipolar re-referencing. The locations and the quantities of the implanted electrodes vary for each patient as the electrodes were implanted based on the needs of the surgical procedures.

The Tables 1.1, 1.2 and 1.3 below show the brain regions and the corresponding number of electrodes implanted for each of the patients respectively. The tables also include the conventional international abbreviations of the brain region names which are used in the figures in later sections of this research paper. The Figure 1.2 shows the electrode placements in the brain for each of the subjects. Generally, most of the electrodes for Subject 1 were implanted in the temporal lobe. For Subject 2 the most number of electrodes were placed in the frontal lobe. Lastly, the majority of Subject 3's electrodes were from the occipital lobe.

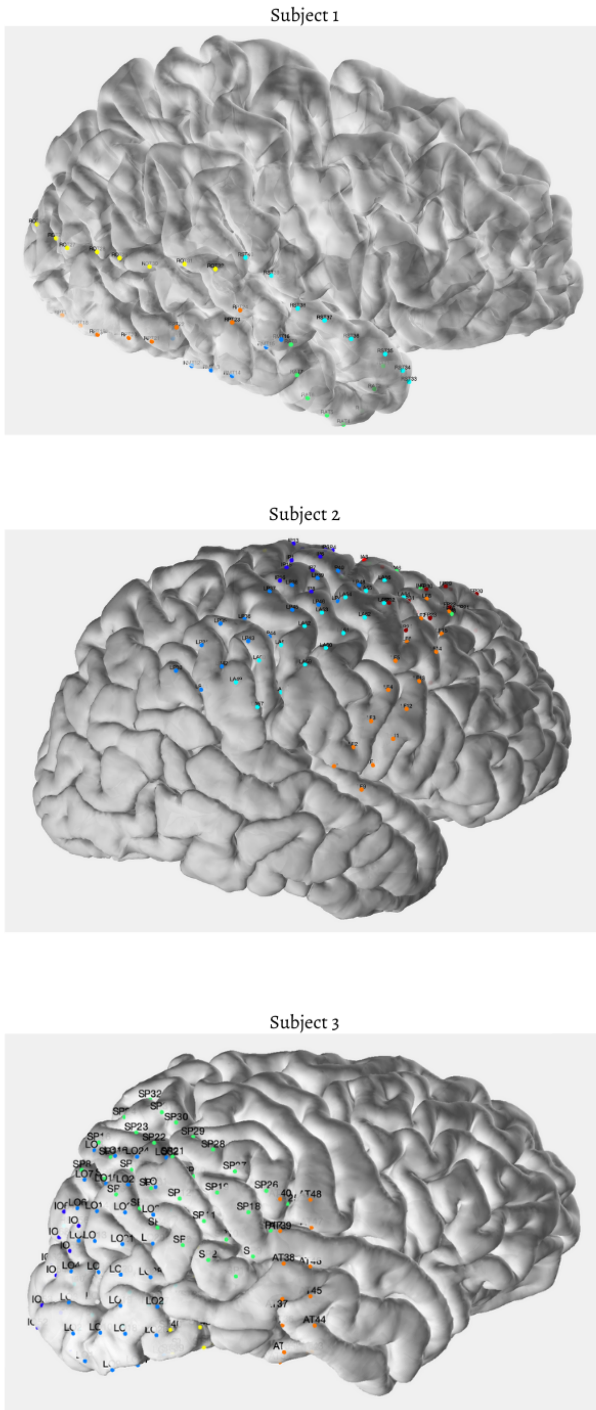

Figure 1.2: This figure was produced with MATLAB and shows the electrode locations in the brain for each of the subjects.

Table 1.1: Number of Electrode Placements in Each Brain Region for Subject 1

Brain Region	International Abbreviation	#Electrodes
Superior Temporal Gyrus - Lateral	G_temp_sup-Lateral	7
Temporal Pole	Pole_temporal	6
Middle Temporal Gyrus	G_temporal_middle	5
Middle Occipital Gyrus	G_occipital_middle	4
Inferior Temporal Gyrus	G_temporal_inf	3
Inferior Occipital Gyrus and Sulcus	G_and_S_occipital_inf	3
Inferior Parietal Lobe - Supramarginal Gyrus	G_pariet_inf-Supramar	3
Inferior Frontal Gyrus - Opercular	G_front_inf-Opercular	2
Inferior Parietal Lobe - Angular Gyrus	G_pariet_inf-Angular	2
Fusiform Gyrus (Lateral Occipitotemporal Gyrus)	G_oc-temp_lat-fusifor	2
Subcentral Gyrus and Sulcus	G_and_S_subcentral	1
Occipital Pole	Pole_occipital	1
Lingual Gyrus (medial Occipitotemporal gyrus)	G_oc-temp_med-Lingual	1

Table 1.2: Number of Electrode Placements in Each Brain Region for Subject 2

Brain Region	International Abbreviation	#Electrodes
Superior Frontal Gyrus	G_front_sup	37
Middle Frontal Gyrus	G_front_middle	13
Precentral Gyrus	G_precentral	12
Postcentral Gyrus	G_postcentral	8
Inferior Parietal Lobe - Supramarginal Gyrus	G_pariet_inf-Supramar	6
Anterior Cingulate Cortex	G_and_S_cingul-Ant	4
Inferior Frontal Gyrus - Opercular	G_front_inf-Opercular	4
Paracentral Gyrus and Sulcus	G_and_S_paracentral	4
Precuneus Gyrus	G_precuneus	4
Inferior Parietal Lobe - Angular Gyrus	G_pariet_inf-Angular	3
Mid-Posterior Cingulate Cortex	G_and_S_cingul-Mid-Post	2
Mid-Anterior Cingulate Cortex	G_and_S_cingul-Mid-Ant	2
Middle Temporal Gyrus	G_temporal_middle	1
Inferior Frontal Gyrus - Triangularis	G_front_inf-Triangul	1
Subcentral Gyrus and Sulcus	G_and_S_subcentral	1
Superior Temporal Gyrus - Lateral	G_temp_sup-Lateral	1
Cuneus Gyrus	G_cuneus	1

Table 1.3: Number of Electrode Placements in Each Brain Region for Subject 3

Brain Region	International Abbreviation	#Electrodes
Lingual Gyrus (medial Occipitotemporal gyrus)	G_oc-temp_med-Lingual	35
Cuneus Gyrus	G_cuneus	21
Inferior Parietal Lobe - Angular Gyrus	G_pariet_inf-Angular	20
Occipital Pole	Pole_occipital	17
Precuneus Gyrus	G_precuneus	10
Inferior Occipital Gyrus and Sulcus	G_and_S_occipital_inf	10
Middle Occipital Gyrus	G_occipital_middle	8
Inferior Parietal Lobe - Supramarginal Gyrus	G_pariet_inf-Supramar	8
Middle Temporal Gyrus	G_temporal_middle	7
Superior Temporal Gyrus - Lateral	G_temp_sup-Lateral	5
Inferior Temporal Gyrus	G_temporal_inf	5
Superior Occipital Gyrus	G_occipital_sup	4
Intraparietal Sulcus	S_intrapariet_and_P_trans	4
Superior Parietal Gyrus	G_parietal_sup	4
Fusiform Gyrus (Lateral Occipitotemporal Gyrus)	G_oc-temp_lat-fusifor	4
Subcentral Gyrus and Sulcus	G_and_S_subcentral	3
Superior Occipital Gyrus	S_oc_sup_and_transversal	2
Collateral Sulcus	S_collat_transv_ant	2
Superior Temporal Sulcus	S_temporal_sup	2
Unknown	Unknown	1
Lateral Occipitotemporal Sulcus	S_oc-temp_lat	1
Postcentral Sulcus	S_postcentral	1
Lingual Sulcus (medial Occipitotemporal sulcus)	S_oc-temp_med_and_Lingual	1
Superior Temporal Gyrus - Planum Temporale	G_temp_sup-Plan_tempo	1

While the patients were undergoing the ECoG procedure, they were also being video recorded meaning that both their physiological behaviors and neuronal recording information could be synced and analyzed together. The audio-visual recordings have been manually annotated with a millisecond (ms) precision level using Neuro NatusWorks software and matched with the neuronal recordings as shown in the Figure 1.3.

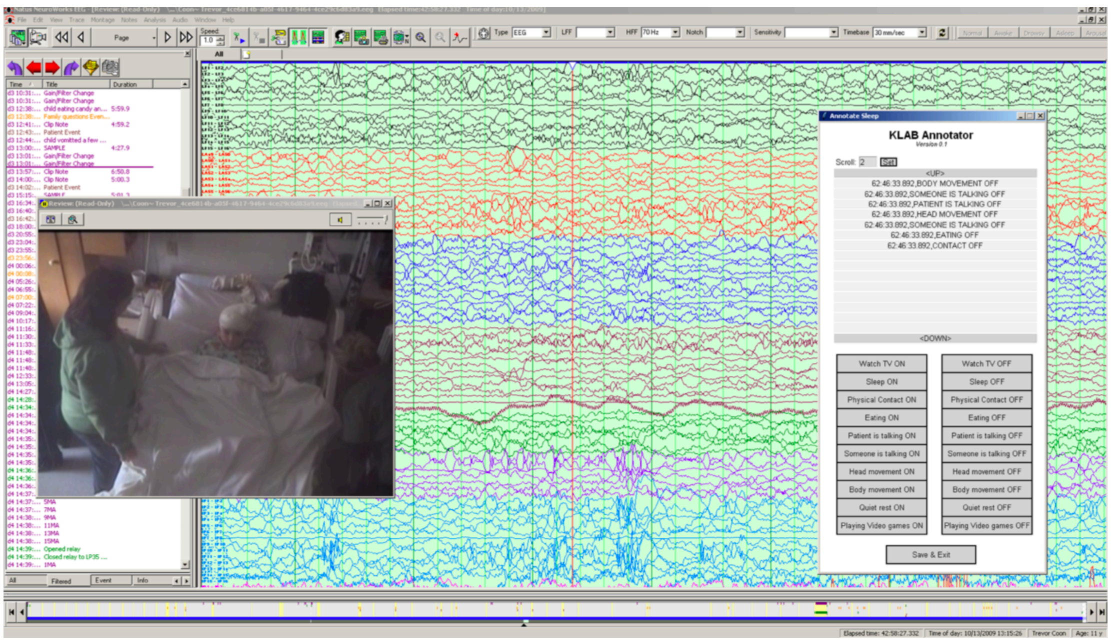

Figure 1.3: The figure is a screenshot from the process of producing the annotations with Neuro NatusWorks showing how the patient's video was superimposed with the brain activity.

The annotated behaviors are:

- **Body Movement**: The patient is moving. This includes any movement but a head movement which is a separate annotation on its own.
- **Arm Movement**: The patient is moving their arms. Both left and right arm movements are captured with this annotation.

- **Leg Movement**: The patient is moving their legs. Both left and right leg movements are captured with this annotation.
- **Head Movement**: The patient is moving their head and neck. The facial movements are not included.
- **Contact**: The patient comes in contact with someone else. The annotation covers all the physical contacts with another human regardless of who initiated it.
- **Someone is Talking**: Someone else in the room is talking other than the patient. The patient does not necessarily need to engage in the conversation.
- **Patient is Talking**: The patient is engaging in a conversation or makes requests.
- **Sleep**: The patient is sleeping or relaxing with their eyes closed.
- **Quiet**: There is no noise in the room and the patient is typically not engaging in any activities.
- **Eating**: The patient is actively consuming food or drinking.
- **Video Games**: The patient is playing video games on the computer.
- **Watching** TV: The patient is actively watching the TV.

Please note that even though the majority of the annotated behaviors overlap across the patients, not all of the patients have the same set of annotated behaviors. The Table 1.4 demonstrates what the exact annotated behaviors are for the three subjects analyzed in this research paper.

Table 1.4: The annotated behavior distributions among the subjects

	Subject 1	Subject 2	Subject 3
Body Movement		Yes	Yes
Arm Movement	Yes		
Leg Movement	Yes		
Head Movement	Yes	Yes	Yes
Contact	Yes	Yes	Yes
Someone is Talking	Yes	Yes	Yes
Patient is Talking	Yes	Yes	Yes
Sleep	Yes	Yes	Yes
Quiet		Yes	Yes
Eating	Yes	Yes	Yes
Video Games	Yes	Yes	Yes
Watching TV	Yes	Yes	Yes

All the behaviors were annotated in a linear fashion with binary ON and OFF labels. For example, once the patient would start eating the annotations for EATING behavior would be marked ON for each sample of neuronal recording until the patient would stop eating and the annotations would become OFF. To summarize, the neuronal recordings make a tabular X_i with $m_i x n_i$ where $i \in \{1, 2, 3\}$ is the subscript corresponding to the subject, m_i is the number of samples and n_i is the total number of electrode channels for the i_{th} subject. The annotations make a tabular Y_i matrix with $k_i x 10$, where k_i is the number of total annotated samples and 10 is the number of annotated behaviors for the i_{th} subject. The details about the ECoG data for each of the subjects is provided in the Table 1.5 which contains information about the patient's gender, age, total number of electrodes, total number of bipolar electrodes, the total number of available ECoG samples and the sampling rate.

Table 1.5: Details about the ECoG data received from the three subjects.

Subject	Gender	Age	#Electrodes	#Bipolar Electrodes	Total Samples	Sampling Rate
1	Male	32	40	35	82235127	250
2	Male	11	104	91	134834457	250
3	Female	16	176	154	12889503	256

1.2 Feature Engineering

Before performing any classification tasks the data have been carefully processed and the features have been engineered to decode meaningful information from the time-series of signals. The general idea was to divide the time series into time windows/intervals and calculate an aggregative metric corresponding to the signal in that time window, e.g. binning one second of iEEG samples in one window and calculating the maximum amplitude. Firstly, this would achieve aggregation of the data from a millisecond precision to a second precision to alleviate human reaction errors while annotating. Secondly, the feature engineering process allows us to denoise the data and incorporate domain specific knowledge on time-series, signal processing and neuroscience. Lastly, this approach makes the data computationally more efficient to process.

A few methodologies of feature engineering have been implemented to present different approaches and sets of assumptions, specifically, about how to choose the windows which will divide the neuronal recordings with respect to behavioral annotations' ON and OFF intervals. Furthermore, both bipolar and original referencing were considered resulting in fours different sets of features.

The main approach with the window selection was to create one-second windows which, depending on the sampling rate, correspond to either 250 samples per second for Subject 1 and 2, and 256 samples for Subject 3. As a classification sensitivity analysis two-second windows were also calculated for the best set of features and the classification results can be found in the results section.

1.2.1 Equally Spaced Windows

The main question for implementation was deciding how the starts of the windows should be chosen given the time series of neural recordings and the behavior annotations. The easy, more obvious, technique is to divide the entirety of the series into equally spaced one-second intervals. The diagram below (Figure 1.4) shows how the signal which has corresponding consecutive periods of ON and OFF labels for the behavior annotations is divided into equally spaced one-second windows. For the purposes of reducing the ambiguity in the classification, the windows which had samples from both ON and OFF annotations were discarded, even though one might argue that the most interesting information is contained at the change points of the behaviors.

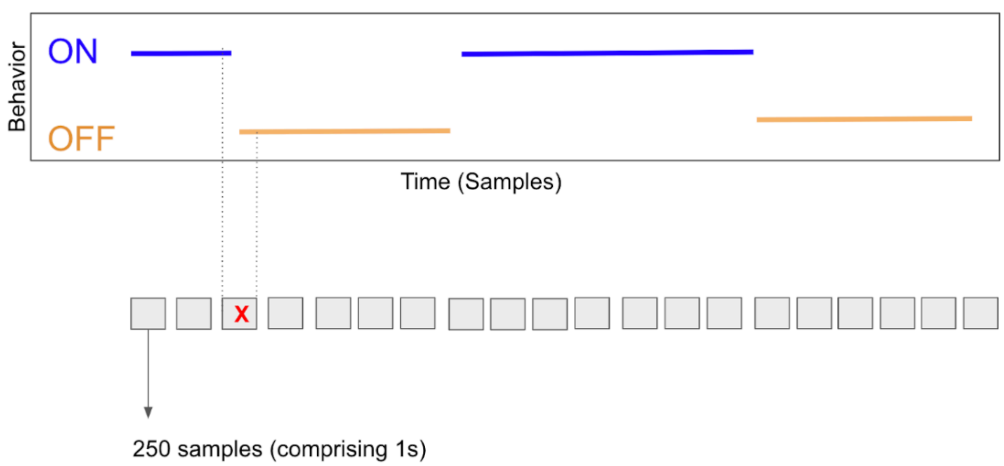

Figure 1.4: The graphic portrays the methodology of dividing the signal into equally spaced windows for feature engineering. The blue lines represent the part of neuronal signals which were annotated ON as the given behavior and the orange line represents the consecutive OFF parts of the signal.The grey squares are the equal time windows.

1.2.2 Windows Adjusted to the Start of Behavior Change

As discussed in the previous approach, the discarded windows which are at the intersection of behavior changes - e.g. the second which contains the state of the person starting to consume food - might contain valuable information which is being lost with the initial approach. Hence, an alternative approach was developed to adjust for the start of behavior changes. In this approach, the starts of window intervals reset with the start of ON or OFF annotation periods. For example, the one-second windows will start when the person starts eating and will be equally spaced until the person stops eating. Then the window start will be reset to match the exact second when the person stopped eating and will continue until the person will start eating again, for which the start will be reset again to match the new behavior change. The neuronal samples at the end of each behavior ON or OFF annotation interval will be discarded if they do not make a full one second window. This methodology is visually presented in the Figure 1.5.

Since this approach is very behavior specific, each behavior ends up having its own separate one-second windows of neuronal samples and, hence, the downside of this method is that it requires to compute different sets of features for each behavior and is computationally very costly.

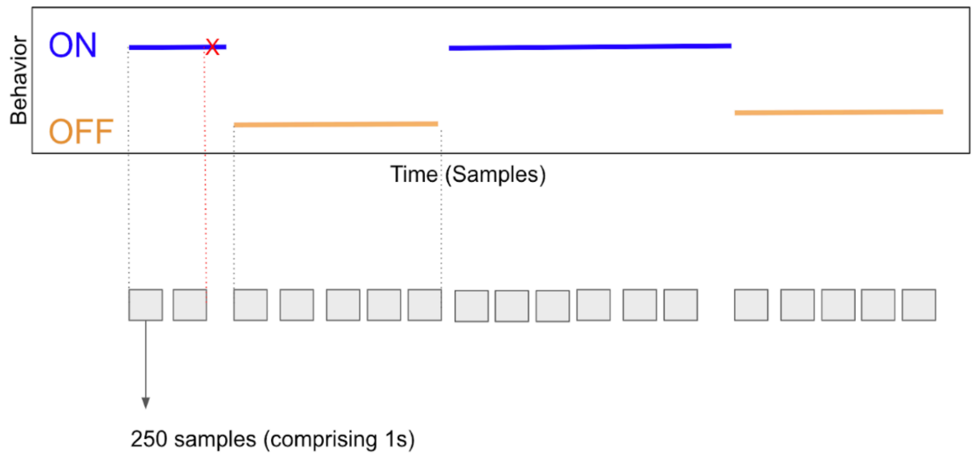

Figure 1.5: The graphic portrays the methodology of adjusting the start of windows selection to the beginning of the behavior change. The blue lines represent the part of neuronal signals which were annotated ON as the given behavior and the orange line represents the consecutive OFF parts of the signal.The grey squares are the equal time windows.

1.2.3 Referencing

A common way to process the iEEG/EEG electrode voltages is by introducing a reference montage. Using a reference montage means to get the potential difference between the electrode and a reference point which will ideally be a static reference value but often is another electrode. The most common reference method is bipolar montaging. The bipolar reference is when the potential difference is calculated from the pair of electrodes where one serves as a reference for another. In my case, I used the neighboring electrode pairs for bipolar re-referencing, as the information about the electrode locations is available. Bipolar reference is defined as $V_{ij}^{bip} = V_i - V_j$, where V_i and V_j are the voltage signals from i_{th} and j_{th} neighboring electrode pairs at a given time. This allowed me to capture the local changes in the voltage potentials, canceling out the signals which are the same throughout the brain areas. However, bipolar re-referencing might result in loss of information, as it only captures the local changes in potential and not global. For this reason and as a case study, I also considered the raw voltage signals in my feature engineering process.

The re-referencing choices combined with the two methodologies of window start selection resulted in four main feature sets which are:

- Methodology A: equally spaced windows with bipolar referencing
- Methodology B: equally spaced windows without bipolar referencing
- Methodology C: windows adjusted to the behavior change with bipolar referencing
- Methodology D: windows adjusted to the behavior change without bipolar referencing.

Please, note, that for the clarity and simplicity of this paper, the main figures in the results section are based on the features derived through Methodology A, except in the subsection where the classification results are reported and compared for the different feature sets.

1.3 Feature Calculation

1.3.1 Time Series Features

After the windows were chosen, the main objective of the feature engineering phase was to choose metrics/features to capture the most relecant information from the channels. Since the data is a time series of voltage samples it made sense to calculate typical times-series features. In my case, I chose to calculate the root mean squared (RMS), mean absolute value (MAV), the minimum amplitude (MIN) and the maximum amplitude (MAX) of the one-second windows. The general overview of what type of information these features capture for the raw signals is presented in Figure 1.6.

- **RMS**

  Root Mean Squared for a window of 250 samples is calculated by squaring all the values, averaging them and then taking the square root of the mean. The equation is shown below where v_i is the individual IFP.

  $$RMS(v) = \sqrt{\frac{1}{n}\sum v_i^2}$$

  RMS is widely used for exploring the time series of electrical signals, especially, when it comes to defining and working with sinusoidal voltages.[11] Since the iEEG data also has an oscillatory nature, RMS values will allow me to capture information about the amplitude of the wave. The various amplitudes of the waves are hypothesized to be associated with different human behaviors which makes it even more of an attractive feature for this study.

- **MAV**

  Mean Absolute Value (MAV) is computed by taking the absolute values of the individual voltages and then taking the average for the window as defined below:

$$MAV(v) = \frac{1}{n}\sum |v_i|$$

The MAV will capture the average values of the voltages without letting the positive and negative values cancel each other out. In this way, for each window I will maintain the average voltage intensity information.

- **MIN**

  Minimum value is computed in a very straightforward way: I take the minimum of the voltages for the given window.

  $$MIN(v) = min(v_i)$$

  I wanted to keep information about all the extreme activities as it will illustrate the drops which may be caused by stimulation. The minimum values as a feature keep track of how deep the lows of the IFPs can get.

- **MAX**

  The purpose of computing the maximum values is very similar to the minimum values, except this feature keeps track of the signal voltage peaks.

  $$MAX(v) = max(v_i)$$

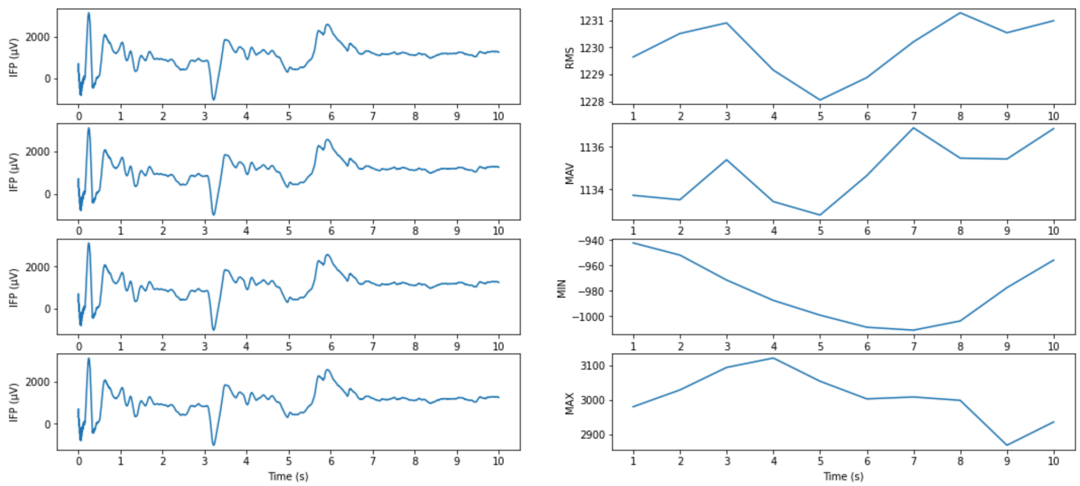

Figure 1.6: The figure compares the original signal with the RMS, MAV, MIN, MAX features received for the 10 second interval.

1.3.2 Frequency Features

ECoG data used in this study has an oscillatory nature and can be thought of as a sinusoidal electrical wave which could be represented in a frequency domain. Neuroscientists commonly use five different frequency bands when working with brain waves: ALPHA, BETA, DELTA, THETA, GAMMA. The research showed that the various frequency bands are associated with different physiological states and behaviors. Hence, naturally, the frequency features become very valuable tools in decoding the behaviors in this research. The detailed description for each of the bands and their calculations follows in the subsections. Moreover, the Figure 1.7 shows the overview of the general information captured from the raw signal by contrasting the features computed with the original signal.

- **DELTA**

  Relaxation, deep and restorative sleep are the key states that come to mind when discussing Delta waves. Delta wave frequencies range from 0.5-3 Hz. These waves are usually prominent among infants' brain activity and those who have brain injuries, inability to think or severe ADHD.[38] The DELTA feature is calculated by approximating the power spectrum with Welch's method and measuring the density over the 0.5-3 Hz range. For DELTA and the following frequency band calculations the spectral power density estimation is carried out with Python's Scipy package's signal.welch method.[42]

- **THETA**

  Theta activity is a normal EEG feature which covers less than 10% of the recording.[13] It spans the frequency range of 3-8Hz. In 2017, Thitz et al showed that the theta-burst stimulation could improve the episodic memory. As mentioned in "Functional Neuromarkers for Psychiatry" by D. Kropotov there is only one theta rhythm in the healthy brain—the frontal midline theta.[23] The presence of theta activity in other parts is deemed to be an abnormality. This type of abnormality is usually seen in autistic patients. THETA feature in this study is calculated by finding the density of frequencies in the range of 3-8Hz from a spectrogram estimated by Welch's approach.

- **ALPHA**

  Alpha waves are those frequencies which span the range from 8 to 12 Hz and usually peak at 10 Hz. Mindfulness, focused relaxation, the ability to mentally coordinate and access to controlled knowledge are states which are linked with Alpha waves as suggested in "Alpha-band oscillations, attention, and controlled access to stored information" by W. Klimesch.[22] Moreover, Alpha frequencies were shown to dominate when the person was resting with eyes open and disappear during the sleep.[29] Alpha waves vanish when the person is performing active tasks. The ALPHA feature in this study is calculated by estimating the spectral power density by Welch's method over the specified time window and then taking the average density of the corresponding frequency range from 8Hz to 12Hz.

- **BETA**

  Beta waves are the high-frequency, low-amplitude waves ranging from 12-30Hz (in some literature it may be defined to reach 40Hz).[30] They are associated with conscious thoughts, logical thinking and are observed mostly when the person is fully awake. Too much of Beta frequencies is shown to cause anxiety and, too little leads to ADHD and poor cognition.[4] Even within the BETA band, the neuroscientists differentiate three subcategories: low beta waves(12-15Hz) of quiet and focused states; mid-range beta waves(15-20Hz) signifying increase in energy and performance; and high beta waves(20-30Hz) showing very high energy, stress and anxiety. The BETA feature of this study was calculated using the Welch periodogram for the one second windows ranging from 12Hz-30Hz.

- **GAMMA**

  Gamma waves are the fastest oscillating waves to be observed in our brain activity. The Gamma band is usually defined to be greater than 30Hz and includes frequencies up to 100Hz. The prominence of fast oscillating waves can lead to anxiety but the suppression can also lead to loss of attention span and learning disabilities. When obtained in normal amounts, the presence of Gamma activity is associated with consciousness and multidimensional activities such as binding of the senses (smell, touch, sight, taste).[4] GAMMA feature, as all the other frequency features, is also calculated using the Welch's method for approximating the spectral density over the time window and getting the density over the specified Gamma range (30-100Hz). Please note that the frequencies at 60Hz are not taken into account as they were previously Notch filtered to avoid the recording device's electrical oscillations.

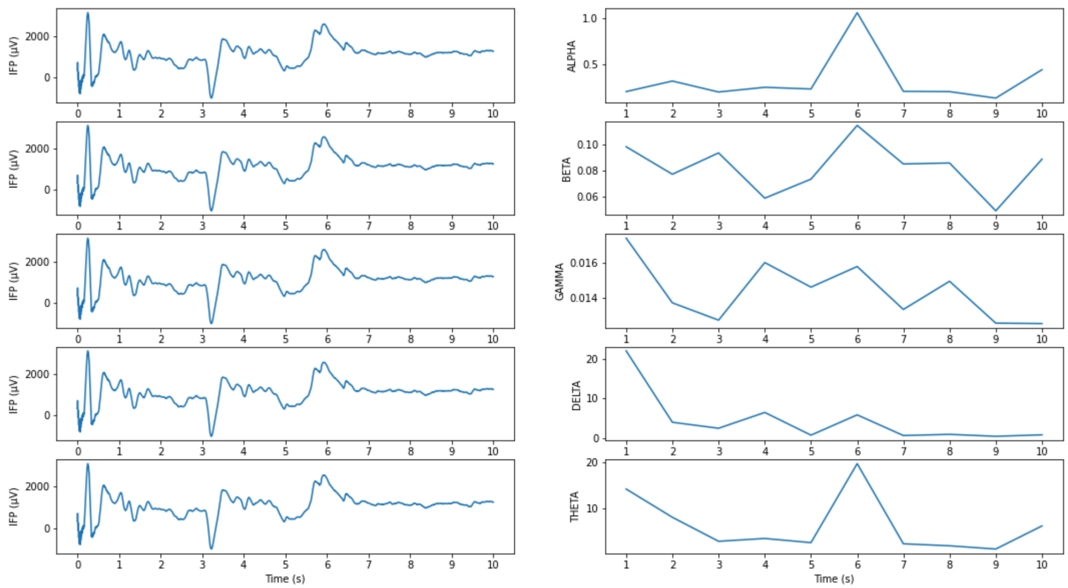

Figure 1.7: The figure compares the original signal with the ALPHA, BETA, GAMMA, DELTA, THETA features received for the 10 second interval.

1.4 Feature Exploration

Before diving into any of the classification results, I wanted to see if my engineered features are of any meaning, so I used dimensional reduction algorithms such as Principal Component Analysis (PCA) and t-distributed Stochastic Neighbor Embedding (t-SNE) for visualizing and interpreting my features in 2D and 3D. Principal Component Analysis is a technique of computing the principal components of the dataset in such a way that the information loss is minimized. [20] In a nutshell, the principal components constitute an orthonormal basis for the data where the different dimensions of data are linearly independent from each other. After the principal component analysis is carried out the principal components are ordered from the highest information variance to the lowest. So for the dimensional reduction purposes the set of the first few principal components usually includes most of the variance of the dataset and can be used to visualize the information in the lower dimensional space.

T-distributed Stochastic Neighbor Embedding (T-SNE) is a nonlinear dimensionality reduction tool where the data points are assigned values in such a way that the similar data points end up close to each other in the reduced dimensional space.[41] Essentially, it assigns a distribution for the data points in the original high dimensional space and another distribution in the specified lower dimensional space, and minimizes the Kullback-Leibler divergence (KL divergence) between these two distributions. Both t-SNE and PCA dimensionality reduction methods were utilized to visualize my features in the lower dimensional space and the figures are reported in the results section. I have used the open source sklearn library in Python for t-SNE(sklearn.manifold.TSNE) and PCA (sklearn.decomposition.PCA).[33]

1.5 Model Selection

In machine learning, the models which learn functions based on already observed and annotated examples of feature-outcome pairs are called supervised learning models. From that perspective the decoding of human physiological behaviors given the neuronal recordings (observed data/features) and video annotations (outcomes) can be formulated as a supervised learning problem. There are many common models used for supervised learning tasks presenting different levels of theoretical complexities from Logistic Regression, k-NN (k Nearest Neighbors) to neural network based models such as CNNs (convolutional neural networks) and attention based transformers. While I tried many models and I reported some of their results in one of the subsections, most of the main results presented in the paper are accomplished by Random Forests. The full list of the models used in this paper is: Logistic Regression, Random Forests, k-NN, SVM with stochastic gradient descent, Gradient Boosting Classifier). All of the classification is performed in Python with the utilities of its sklearn library.[33]

1.5.1 Random Forests and Model Tuning

The objective of this research was not only reaching high classification performance but also be able to interpret the results and formulate hypothesis about the biological drivers of the decoding success for each behavior. Hence, I chose to use Random Forests as a main classifier throughout the paper because it allows to capture non-linear classification boundaries while also providing room for interpetability of feature importances.

Random forests is a decision tree based ensemble learning algorithm used both for classification and regression supervised learning tasks. During the training phase the model constructs multiple bootstrapped decision trees and ensembles the most popular vote for a classification outcome.[8] The different design choices during the training process include but are not limited to the number of estimator trees, the maximum depth of the trees, the criterion of the decision split (Gini impurity index is the most common one) and maximum number of features to be used in the classification for each tree. In order to obtain the best possible classification results, in this paper the random forests of the main results are fine-tuned using a hyperparameter grid search. The main two tuned hyperparameters are the number of estimators and the maximum depth of the trees. Random forests are more efficient computationally and more explainable than the neural networks but still achieve great classification results. With a dataset as complex as this one I prioritized having a model which could be easily tuned and explained. To ensure the stability of the classification results, all of the models in this study are fitted 10 times with 5-fold cross validation.

1.6 Class Imbalance

The training process in classification is very sensitive to the class distribution in the data. Class imbalance in the dataset is the scenario when one of the classes has more observations that the other class and it can lead to an incorrect interpretation of accuracy results. In our case the data is heavily imbalanced for all of the classes. To address this issue, I use a sampling procedure: they either undersample the dominant class or oversample the class which does not have a good representation in the dataset. Sometimes a technique of generating synthetic data is used if we know the data generation distribution. Since the underlying data generating distribution of neuronal signals is not known to humankind, yet, synthetic data generation was not appropriate for this study. On the other hand, the data is big and oversampling would mean adding to the computational costs of obtaining classification results. Hence, for this study purposes I chose to implement undersampling techniques.

1.6.1 Undersampling through Simple Random Sampling

Simple random sampling (SRS) is a probability sampling method where each member of the selection population has an equal chance to be chosen for a sample. Simple random sampling is used widely in experimental research as it is shown to form a sample which is representative to the population - a characteristic which makes the SRS an attractive undersampling option for this paper as well. Since there are 10 behaviors to be classified, a different SRS was undertaken for each of the behaviors, where I would find the number of nondominant class points and sample the exact same number of points from the dominant class. Hence, before each classification procedure I would have perfectly balanced classes.

Please note that all the undersampling techniques suffer from the drawback of possible valuable information loss, and for this study we move forward with this option as the data is big and I hypothesize that the undersampled samples are representative of the population trends.

1.7 Evaluation Metrics

The classification results are only as good as the evaluation metrics show at the end, so it is very important to choose prompt evaluation metrics. The most obvious and common evaluation metric is the classification accuracy, which is defined as the proportion of the sum of true positives(TP) and true negatives(TN) over the number of predictions (N_{pred}).

$$Acc = \frac{(TP + TN)}{N_{pred}}$$

1 or 100% is a perfect accuracy score, while 0.5 or 50% means that the model predicts at random.

While accuracy gives a good measure about the data points which were classified correctly, it lacks the ability to communicate where and why the errors occurred. That is why the data scientists always look into other metrics such as False Positive Rate, True Positive Rate and/or the area under receiver operating characteristic curve (ROC-AUC or simply AUC in this paper). The receiver operating characteristic curve is a plot portraying the relationship between false positive rates and true positive rates as the decision threshold changes. Area Under the Curve (AUC) is a measure of how well the two classes in the binary classification can be distinguished from each other.[7] 1 or 100% is the best AUC score that can be possibly achieved meaning that the two classes can be perfectly distinguished from each other. 0.5 or 50% means that the model is very poor in distinguishing from each other and predicts at random.

Even though, the AUC is not strictly better than the accuracy as a measure, it is still a different and more robust way of evaluating the classifier, as it does not directly apply to a single decision threshold, but shows how robust the classifier would be in differentiating between the classes even when the threshold varies incrementally. It also shows the probability of the randomly chosen ON example being ranked above the OFF example. I report both Accuracy and AUC for the main classification results. For the simplicity of the paper, the supplemental figures and analysis are only with the model AUCs.

1.8 Sensitivity and Control Analysis

Since the research problem involves very challenging settings and noisy data, it is important to run model sensitivity and control analysis with different limiting considerations and ensure that the results are significant and not by chance. The controls discussed in this section are the shuffle test (1.9.1), priority adjacent ON and OFF sampling(1.9.2), time invariance analysis(1.9.3), feature engineering design choices(1.9.4), window selection size (1.9.5).

1.8.1 Shuffle Test

In this study, the shuffle test is defined to be the process where the classification model is fitted to the randomly shuffled annotations. This test makes sure that when the annotations are randomly shuffled the classification model does not decode the behaviors well. The expectation for the classification model is to have around 50% accuracy/AUC, as this model should be only as good as the random prediction generator. Performing a successful shuffle test ensures that the real classification results are not due to random chance and that there are no obvious underlying biases detectable by this test.

1.8.2 Priority Adjacent ON and OFF Sampling

In the previous sections, I have discussed the importance of class imbalance and stated that as a main technique to correct the class imbalance I used an undersampling technique through a simple random sampling. While simple random sampling has many advantages, in this study it may lack the ability to capture the most interesting and sensitive data points: the adjacent ON and OFF points. It is a valuable question to consider whether the classifier is able to correctly decide whether the behavior has changed at the very first second of the behavior change. For example, can I correctly classify that the person is speaking the moment they started uttering the words? To answer this question, I have implemented a custom sampling algorithm which is prioritizing the adjacent ON and OFF pairs while undersampling. The algorithm is implemented through a priority queue, where the priority is defined to be the minimum distance between the dominant class (the class to be sampled from) and the closest nondominant class representative point. Using this priority queue the samples which are directly adjacent to the non dominant behavior have the highest chance of appearing in the sample; and the samples which are the farthest away from the behavior change moment have the lowest probability to appear in the sample.

1.8.3 Time Invariance Analysis

One might think that the behaviors over time change meaning that the events annotated as part of the same behavior actually have different representations in the brain as the time progresses. For example, eating lunch in the afternoon is not the same as eating dinner in the evening or watching TV at 3pm might be different from watching TV at 9pm. Not only the activities and environmental confounding variables might be different (the meals are distinct with their sensory palettes), but also there is a sense of time progression that needs to be taken into account in our sensitivity analysis.
To control for this case, the ON and their adjacent OFF events have been divided into train and test groups in such a way, that the training group includes the activities only from the first part of the day and the test group spans the last part of the day. In this way, the models will be trained with the information from the activities at an earlier times of the day and tested with the later ones. This test will uncover if any of our hypothesis about inhomogeneous event distributions within the annotated behaviors is close to reality as in this time-dependent division the test set will have the chance to act as an out of distribution sample for the train.

1.8.4 Sensitivity of Feature Engineering Design Choices

The feature engineering process requires many design choices. In the earlier sections, I described the 4 sets of features I have calculated for this study. The main design choices discussed were the selection of reference montages and the beginning and ends of the one second windows. While most of the results are presented using only Methodology A features, the fourth part of my sensitivity analysis presents classification results with all the feature sets. The purpose of this sensitivity analysis is to ensure that the decoded results are somewhat robust among the different design choices and avoid the notion of "if you torture your data long enough it will confess". With this test I expect to see robustness among all the feature sets, because only in that case I can be sure that there was a true separability in my classification dataset.

1.8.5 Window Selection Size

The features in this study are mainly engineered around 250 samples per second window. While one second window is an easy conventional choice for interpretation and for precision, an interesting question stands on how robust the classification would be if 500 samples were selected per window. For this sensitivity test, I perform classifications with features computed with 500 samples per window scheme and compare the results with the main classification results in the paper.

There's an ancient connection between movement and music. Most languages don't make a distinction between the words 'music' and 'dance.' And we can see that in the brain. When people are lying perfectly still but listening to music, the neurons in the motor cortex are firing.

Daniel Levitin

2 Results

2.1 Annotations

Before presenting any of the classification results it is important to have a look at the data which will be used in the classification. The Figure 2.1 shows the behaviors over time as they interchange between being ON and OFF. For Subject 1 and Subject 2 there are roughly 24 hours of annotated behaviors, while for Subject 3 I have around 13 hours of annotated behavior. The key statistical summary tables (Tables 2.1, 2.2, 2.3) contain key information about the behaviors, specifically the ON events, and their time distributions. The event here and for the rest of the paper is defined to be a period of activity which was uninterruptedly ON or OFF and the event changes when the behavior changes. Both from the table and the figure you can see that SLEEP, QUIET, WATCH TV can be characterized by a fewer but longer events, while behaviors like BODY MOVEMENT, LEG MOVEMENT, ARM MOVEMENT and HEAD MOVEMENT happen more frequently but last shorter. The behaviors like EATING or VIDEO GAMES have a low number of events throughout the annotation period.

Table 2.1: The descriptive statistics of ON Events for Subject 1

Behavior	Mean Time(s)	Median Time(s)	SD Time(s)	#ON Events
ARM MOVEMENT	21.7	6.8	90.6	447
CONTACT	20.7	10.7	26.6	43
EATING	103.6	3.3	217.6	20
HEAD MOVEMENT	25.4	4.7	111.8	343
LEG MOVEMENT	7.8	4.7	12.3	153
PATIENT IS TALKING	81.2	20.8	230.0	211
SLEEP	8939.4	9993.5	4073.3	4
SOMEONE IS TALKING	170.7	30.3	402.9	122
VIDEO GAMES	449.8	470.2	147.7	3
WATCH TV	2045.9	1279.0	2296.5	8

Table 2.2: The descriptive statistics of ON Events for Subject 2

Behavior	Mean Time(s)	Median Time(s)	SD Time(s)	#ON Events
BODY MOVEMENT	13.0	6.8	21.4	1025
CONTACT	24.8	6.4	65.7	239
EATING	37.1	16.4	44.4	27
HEAD MOVEMENT	7.9	3.5	22.8	1032
PATIENT IS TALKING	8.3	2.7	60.8	388
QUIET	1545.9	1036.0	1714.0	10
SLEEP	1583.5	1638.3	946.7	16
SOMEONE IS TALKING	33.2	8.9	100.4	679
VIDEO GAMES	925.6	873.3	481.6	7
WATCH TV	1124.7	605.6	1289.3	19

Table 2.3: The descriptive statistics of ON Events for Subject 3

Behavior	Mean Time(s)	Median Time(s)	SD Time(s)	#ON Events
BODY MOVEMENT	41.3	10.5	112.2	290
CONTACT	104.8	20.2	186.6	90
EATING	232.3	63.0	236.3	13
HEAD MOVEMENT	6.3	3.5	7.1	59
PATIENT IS TALKING	5.1	2.7	6.6	399
QUIET	1268.8	1062.1	1010.2	6
SLEEP	1886.6	1886.6	1171.6	2
SOMEONE IS TALKING	66.5	9.1	845.1	514
VIDEO GAMES	86.6	87.3	37.3	5
WATCH TV	1903.9	1903.9	0.0	1

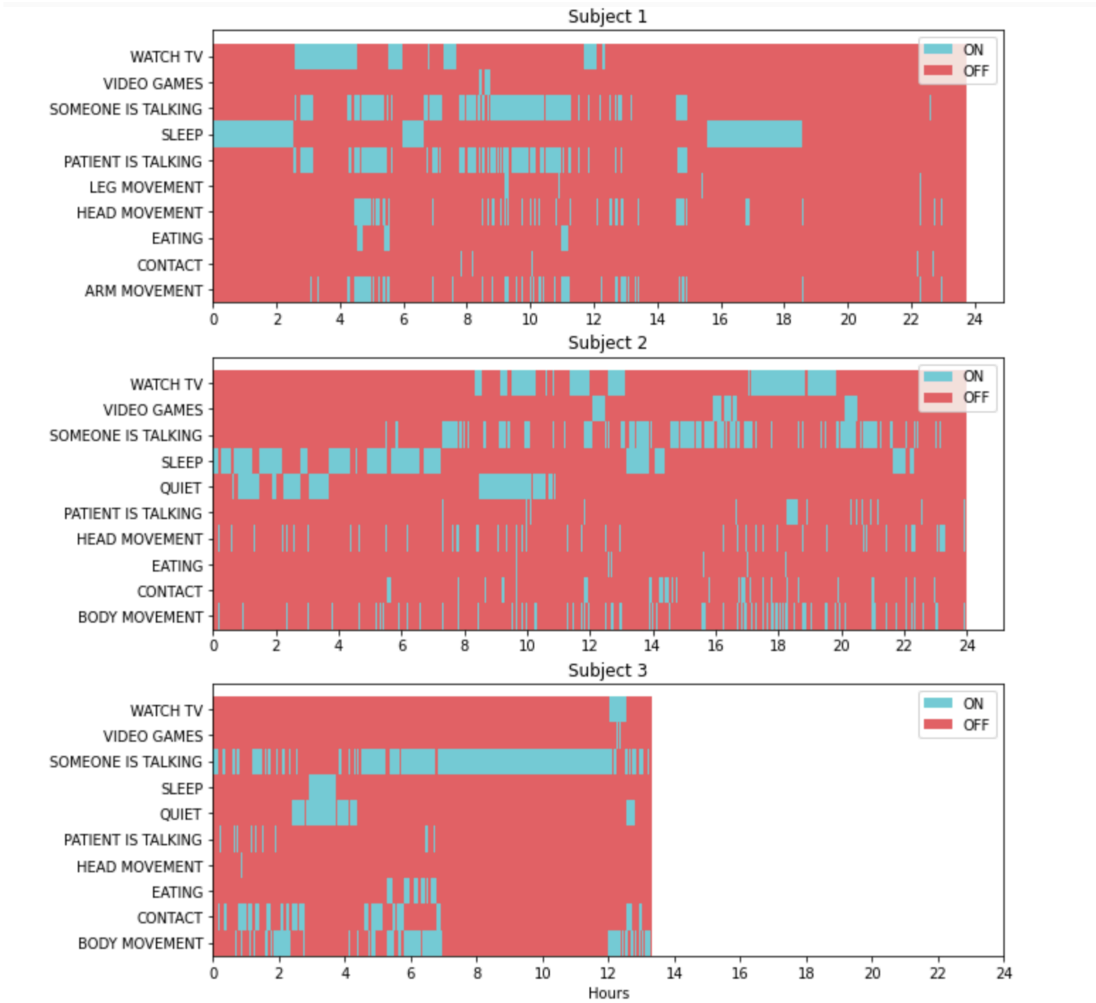

Figure 2.1: The figure illustrates how the events of various physiological behaviors vary over time for the three subjects. The blue color-codes the ON events, while the red shows the OFF events.

Having a look at the overall ON and OFF class distributions for the behaviors on the Figure 2.2 it is clear why the undersampling balancing techniques are essential before performing the classifications. As presented in the figure, the classes for all of the subjects are heavily imbalanced. In some of the cases, such as CONTACT, LEG MOVEMENT, VIDEO GAMES for the Subject 1; EATING and PATIENT IS TALKING for Subject 2; HEAD MOVEMENT and VIDEO GAMES for the Subject 3; the number of ON events compared to the OFF events is so drastically small that my classifier could reach 95% accuracy and better just by classifying OFF for all the points.

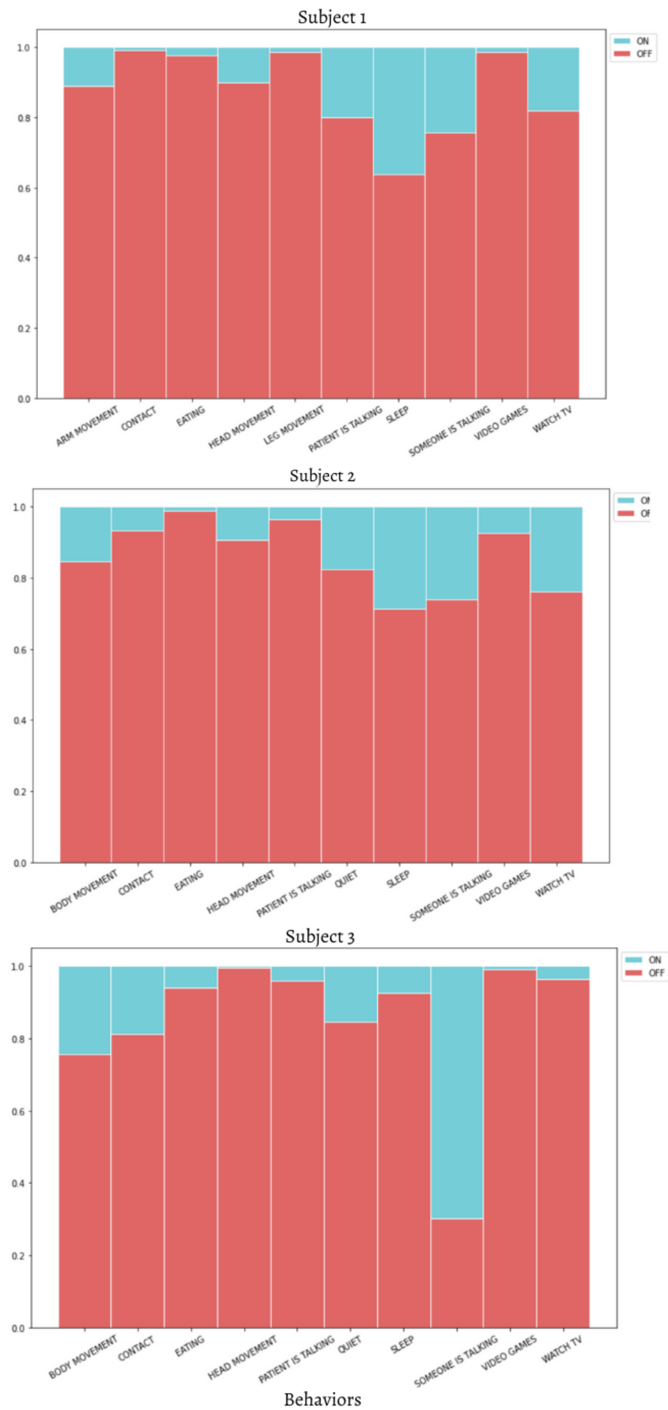

Figure 2.2: The figure illustrates the class imbalance for all of the physiological behaviors for the three subjects. Please note that the y axis is normalized to show the proportions instead of the counts.

2.2 Features

The entire process of feature extraction produces 9 distinct feature sets for 35 bipolar electrodes in the case of Subject 1, totalling 315 features; in the case of Subject 2 I have 91 bipolar electrodes, so 9*91 = 819 features; and finally for Subject 3 I have 154 bipolar electrodes and as a result 9*154=1386 total features. Please note that these reported numbers correspond to the Methodology A of the feature engineering designs. Methodology A will be the feature set used to produce the main results and all the other feature sets will serve as controls in Section 1.8.

The main question that I was interested in concerning my features was how well they show separability among the classes. Below in the Figure 2.3 I show the 3D scatterplot of the first three principal components of 91 GAMMA features for Subject 1 color coded by the EATING ON and OFF behavior. The first scatterplot shows the entirety of the three principal component dimensions and is entirely red, meaning that the ON points can not be distinguished in the transformed space - the red colorcodes the OFF points. However, after zooming in, the right plot shows the ON (blue) points all grouped together in the space. This shows that in the space of GAMMA features EATING behavior is very well defined and there is some logical but nonlinear separability between ON and OFF points.

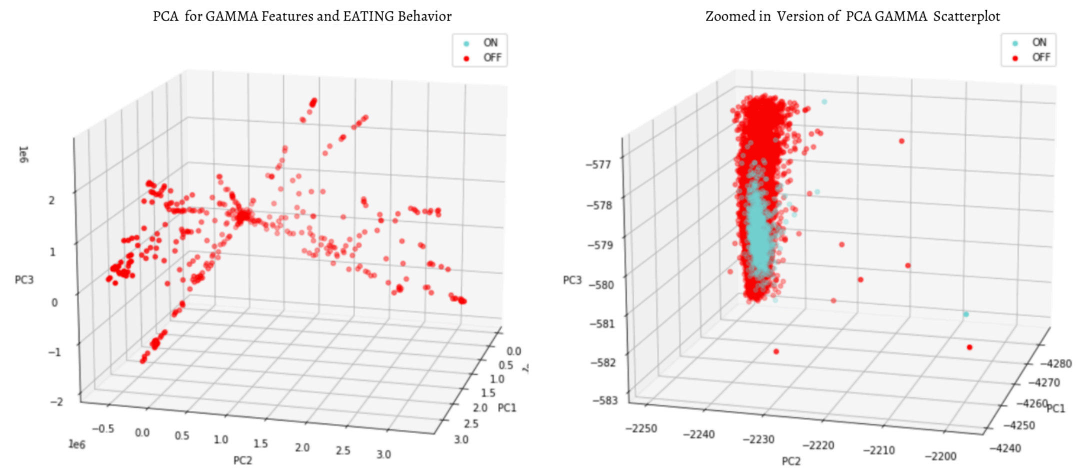

Figure 2.3: The figure illustrates the 3D scatterplot of the first three principal components of GAMMA features' PCA colorcoded for the EATING behavior.

Another example of a good separability is the T-SNE plots of RMS feature sets for Subject 2 SLEEP and SOMEONE IS TALKING behaviors in 2D. For the SLEEP, there is an obvious separation between ON(blue) and OFF(red) points. It is interesting to note that there is one big group of blue points on the right of the plot for SLEEP but there are also smaller groups of blue points towards the left. The annotations for SLEEP were chosen to be marked ON when the person would have their eyes closed since there is no real way of understanding if the person is asleep or just resting eyes closed from the video recording. However, in neuroscience the deep sleep state is different from the eyes closed relaxation, so the smaller groups of ON points could be the results of the mentioned imprecision in annotations.

The plot on the right is the 2D T-SNE dimensionality reduction of RMS features for SOMEONE IS TALKING behavior for Subject 2. From the first sight it might not be apparent but there is an interesting nonlinear separation between ON(blue) and OFF(red) points in the plot. Looking closely, one might notice that all the red points are grouped together and the blue points are clustered at the edges of the red points. While the SLEEP behavior has very well defined ON and OFF states, SOMEONE IS TALKING is not as well-defined and, in fact, many confounding variables may happen in parallel. For example, while someone is talking in the room, the TV might be on or the person might be eating. The fact that there is a notable pattern in the T-SNE distribution and the blue points are still clustered together is remarkable and signals the possibility of successful decoding of the behavior.

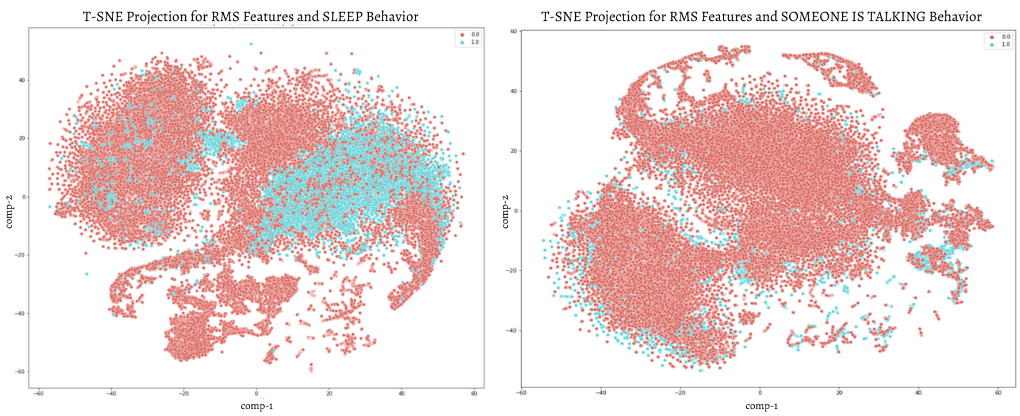

Figure 2.4: The figures illustrate the 2D scatterplots of RMS features' T-SNE projections colorcoded for the SLEEP (on the left) and SOMEONE IS TALKING (on the right) behaviors.

Lastly, I wanted to see how unique the information captured by each of the features was. For this reason I have calculated the pairwise correlation scores for all the features averaged over the number of channels. From the correlation plots one can see that for all of the subjects the RMS and MAV are highly correlated, which makes sense, as mathematically they both get some variation of the average voltage value from the signal in the defined window. MIN features are consistently negatively correlated with RMS, MAV and with the frequency features, even though, for the frequency features the correlation strength varies from patient to patient. For subject 3, RMS features have considerably stronger correlation with DELTA and THETA features. The majority of frequency features have positive correlations among each other.

Subject 1

	RMS	MAV	MAX	MIN	DELTA	THETA	ALPHA	BETA	GAMMA
RMS									
MAV	0.993318								
MAX	0.321071	0.262368							
MIN	-0.578893	-0.534367	0.064382						
DELTA	0.393559	0.346205	0.432936	-0.398825					
THETA	0.241722	0.201886	0.383788	-0.286785	0.708605				
ALPHA	0.174303	0.146237	0.301254	-0.204649	0.607773	0.708577			
BETA	0.179558	0.149646	0.319297	-0.211109	0.562898	0.710646	0.899861		
GAMMA	0.146797	0.121307	0.287668	-0.173624	0.436062	0.643929	0.725356	0.774703	

Subject 2

	RMS	MAV	MAX	MIN	DELTA	THETA	ALPHA	BETA	GAMMA
RMS									
MAV	0.998289								
MAX	-0.858659	-0.871940							
MIN	-0.924950	-0.925287	0.959290						
DELTA	0.148928	0.120709	0.071112	-0.077291					
THETA	0.122559	0.100427	0.058238	-0.067702	0.604648				
ALPHA	0.189793	0.163672	0.088724	-0.092631	0.560594	0.623870			
BETA	0.185407	0.161526	0.086073	-0.091397	0.329275	0.459993	0.813483		
GAMMA	0.170077	0.156183	0.076229	-0.065444	0.055094	0.146424	0.609177	0.749264	

Subject 3

	RMS	MAV	MAX	MIN	DELTA	THETA	ALPHA	BETA	GAMMA
RMS									
MAV	0.988683								
MAX	0.825948	0.797831							
MIN	-0.663796	-0.622714	-0.351037						
DELTA	0.749642	0.751697	0.583514	-0.550207					
THETA	0.609415	0.576141	0.554964	-0.537687	0.543306				
ALPHA	0.392655	0.364086	0.436195	-0.329936	0.227327	0.495880			
BETA	0.257141	0.234694	0.303008	-0.245602	0.097990	0.356560	0.509384		
GAMMA	-0.152703	-0.150672	-0.103079	0.103090	-0.143919	-0.086950	-0.001990	0.118182	

Figure 2.5: The figure shows the average correlations among features for the three subjects.

2.3 Main Results

Taking into account the complexity and noise of the dataset my engineered features reach very impressive classification results in decoding the ten physiological behaviors for the three subjects. Following I present the accuracies(Figure 2.7) and AUCs(Figure 2.8) for all the subjects and each of the behaviors. All of the results in the figures are obtained after equalizing the classes via simple random sampling and fitting a Random Forests model which were fine-tuned for number of estimators and maximum tree depth. After the hyperparameter search I found that the optimal maximum depth to be used in the model was 20 and the corresponding optimal number of estimators was 600 visually presented in Figure 2.6. Given the continuous essence of the features, it makes sense why the deeper trees maximize the accuracy, and given the noisiness of the data it is apparent why the larger number of estimators ensure higher classification performance.

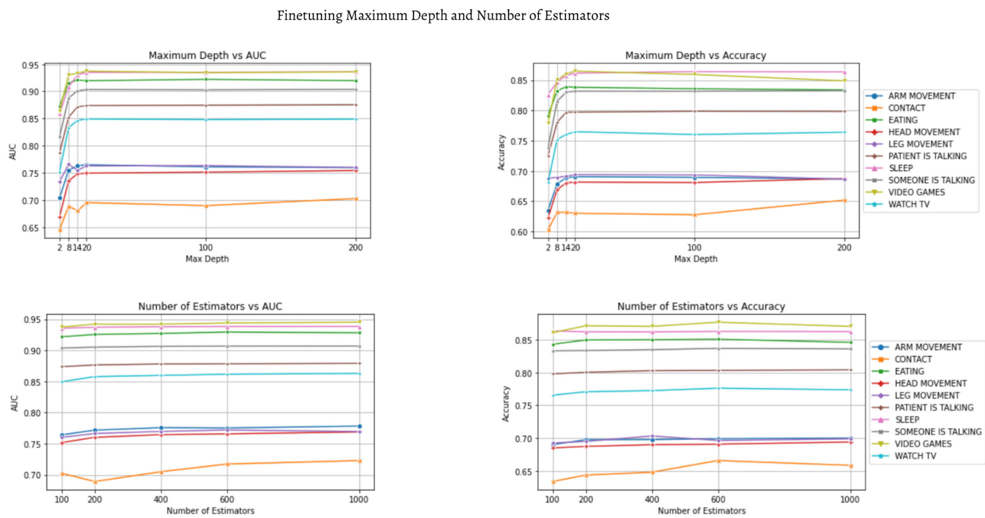

Figure 2.6: The plots show how the maximum depth and number of estimators impact the accuracy and AUC in the classification of behaviors.

From the first sight you can see that EATING and VIDEO GAMES are the two behaviors that are decoded the best ranging from 0.95-0.98 AUCs for all of the subjects, meaning that the classifier is very robust in differentiating between ON and OFF annotations for these behaviors. SLEEP and QUIET are decoded very well for Subject 3 with 0.95 Accuracy and 0.97 AUC. The behaviors which are decoded the worst are the behaviors that involve movement (BODY MOVEMENT, HEAD MOVEMENT, ARM MOVEMENT, LEG MOVEMENT) and CONTACT for Subject 1. Even the aforementioned worst-decoded behaviors reach 0.8 and higher AUCs, showing that the classification models really capture the difference between ON and OFF states.

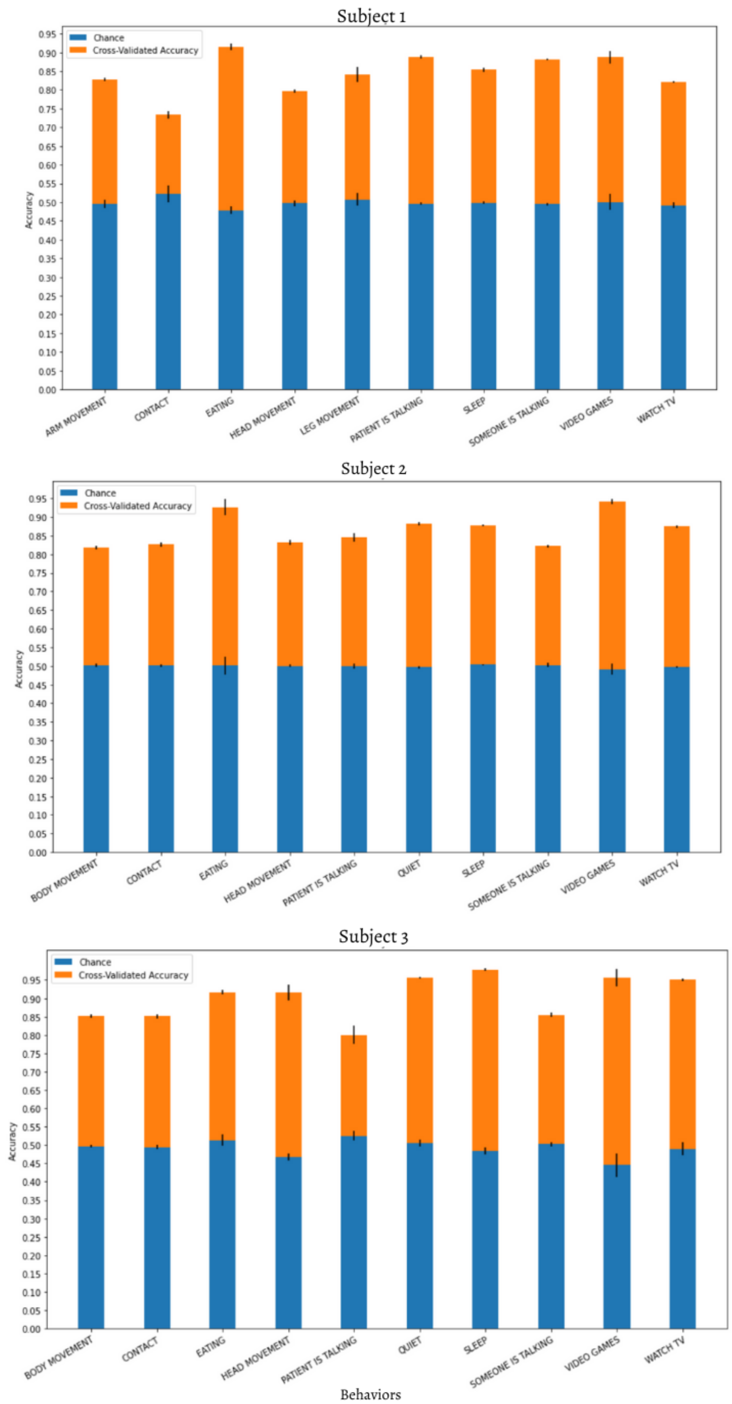

Figure 2.7: The plots showcase the classification accuracies for all the physiological behaviors for each subject. The blue parts of the bars indicate the results received through the shuffle test. The orange part for each bar is the additional cross-validated accuracy.

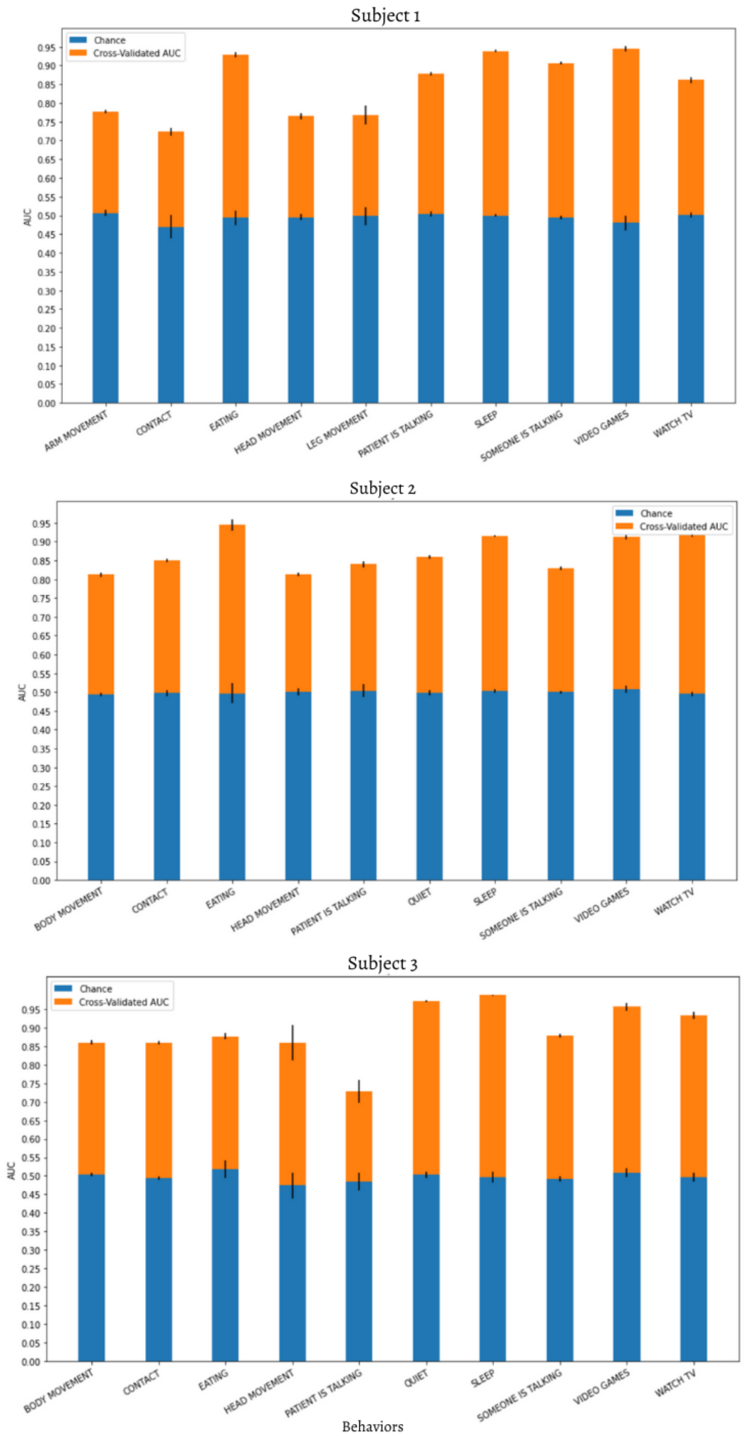

Figure 2.8: The plots showcase the classification AUCs for all the physiological behaviors for each subject. The blue parts of the bars indicate the results received through the shuffle test. The orange part for each bar is the additional cross-validated accuracy.

2.4 Classification Result Comparisons for Various Models

As mentioned in the Section 1.5 , before deciding to choose Random Forests as the main decoding algorithm I have experimented with various algorithms, and in this section I am presenting the results of those experiments. The Figure 2.9 shows the AUC comparison for Logistic Regression, Random Forests, Gradient Boosting Classifier, Support Vector Machines (SVM) with stochastic gradient descent (SGD) and k Nearest Neighbors (k-NN). Consistently, the two worst performing algorithms in the decoding tasks were SVM with SGD and k-NN, and the best two performing classifiers were Logistic Regression and Random Forests. Gradient Boosting Classifier has been exhibiting average results for all the behaviors. It is worthy to note that none of the algorithms had been fine tuned for this experiment and have been run with sklearn's default parameters. While the Logistic Regression and Random Forests have been performing very closely in this experiment, I decided to move forward with Random Forests as the main classifier for this paper as Random Forests would allow me to capture non-linear classification boundaries, which I hypothesized exist in a task as complex as this. Moreover, Random Forests had more design parameters which could be tuned for a better performance than Logistic Regression. And finally, Logistic Regression was requiring at least 10000 maximum iterations to converge with the initial feature set, and it ended up being more computationally costly than the Random Forests.

The overall picture of the behavior decoding has remained the same as presented in the main results section before. EATING and VIDEO GAMES are still the best behaviors to be decoded for all the subjects while CONTACT has the lowest classification performance among all the models.

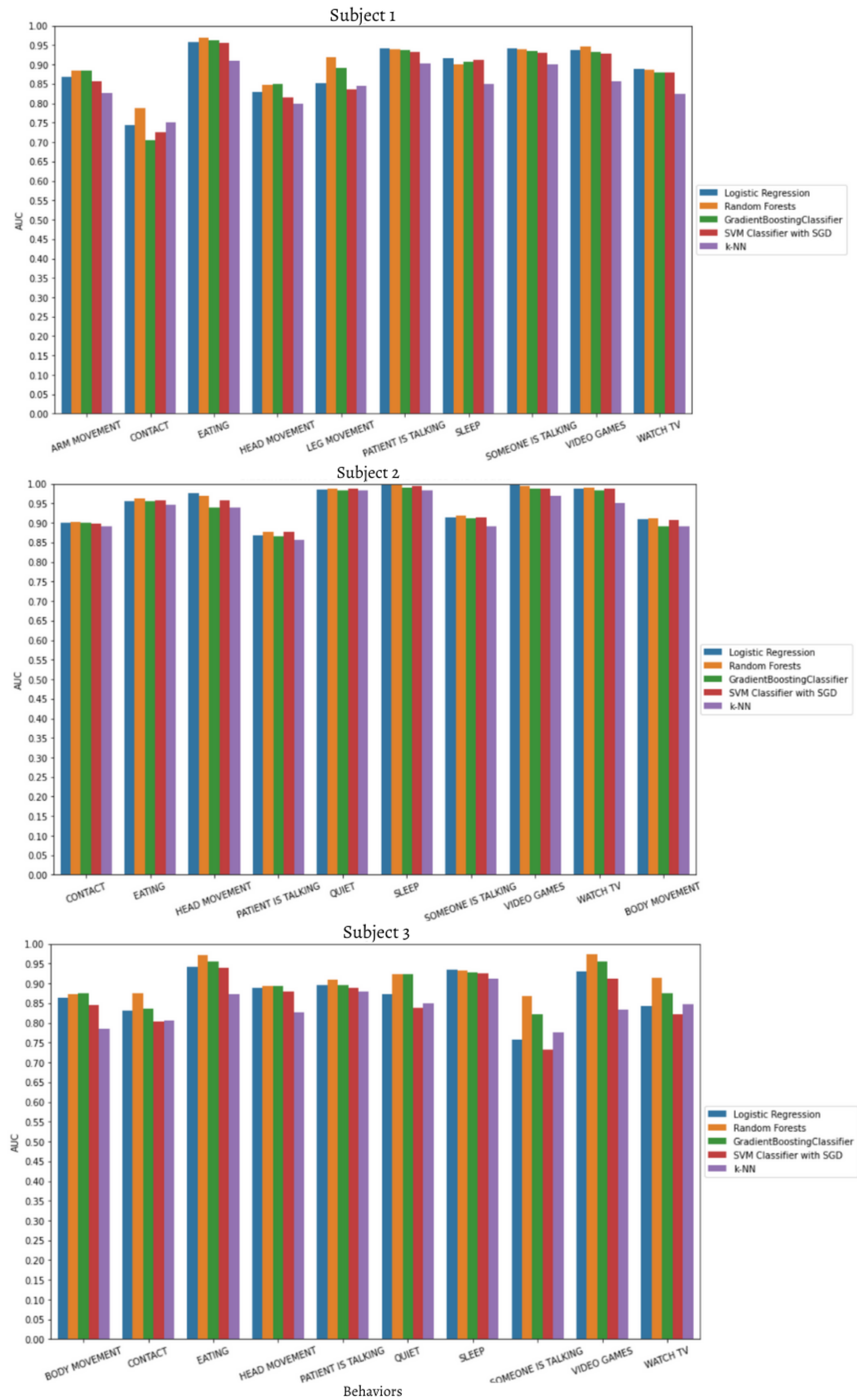

Figure 2.9: The plots showcase the classification AUCs for all the physiological behaviors for various models.

2.5 Classification Comparisons with Time and Frequency Features

Since our feature engineering process involved calculating features based on either the underlying time series characteristics of the voltage waves or the frequency based features, it was natural to run classifications with those two groups of features and see which feature group had the most decoding power. Figure 2.10 below shows the results of these classifications and, unfortunately, there is no easy answer of which feature groups perform better. For Subject 3 the frequency based features perform worse than the time-series features across all the behaviors. For Subject 1 and Subject 2, EATING behavior is decoded better with the frequency-based features. In fact, for Subject 2, the only two behaviors that are decoded better with frequency-based features are EATING and BODY MOVEMENT. All the other behaviors fr Subject 2 are classified better with time series based features. If I were to analyze only Subject 2 and Subject 3 I would somewhat conclude that the time-series based features decode more information than the frequency based features. However, Subject 1 shows a different trend, where most of the behaviors except CONTACT and LEG MOVEMENT are predicted better with frequency-based features. Hence, it is very hard to draw any final conclusions on which feature group is better based on these performances.

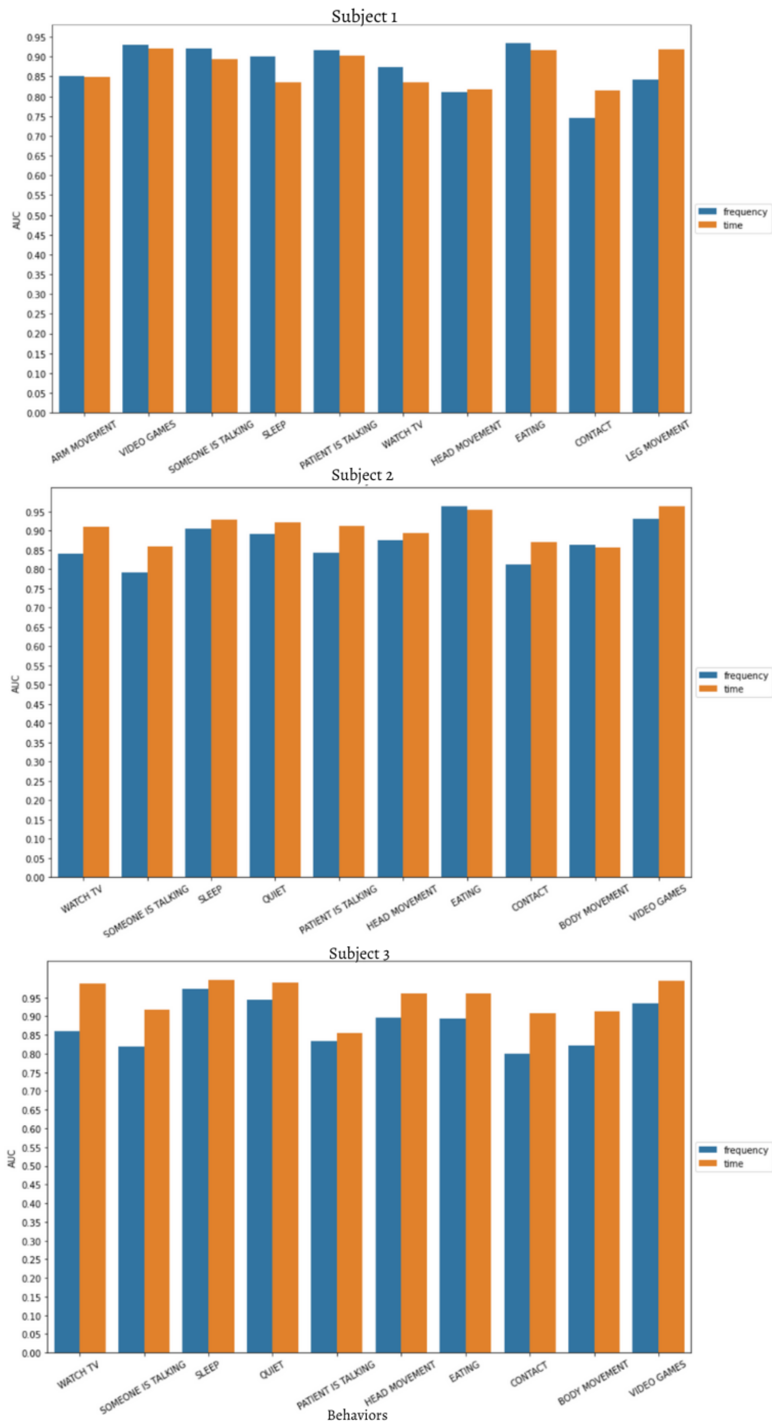

Figure 2.10: The plots showcase the classification AUC comparisons for time vs frequency features among all the physiological behaviors.

2.6 Sensitivity and Control Analysis

2.6.1 Priority Adjacent ON and OFF Sampling

The role of sensitivity analysis is very important in the research as it ensures that the classification performance is not merely due to chance or the result of lucky circumstances. In Section 1.6.1, I discuss that the drawback of the simple random sampling is that it does not necessarily sample the adjacent OFF points to the ON events and, hence, the classification results might be inflated. In this section I present the results of the classification performances after the custom sampling algorithm was applied which prioritized adjacent ON and OFF events as described in Section 1.8.2. The obvious and expected trend is that the classification AUCs for this control are lower than the original classification AUCs in the main figures. One of the contributors to this decrease in performance can be the fact that the annotations of this study are nowhere perfect and the adjacent ON and OFF points might be spuriously annotated. Another logical explanation is that the features (the electrical wave characteristics) do not drastically change at the second the behavior changes and rather flow to the next state gradually. The Figure 2.11 illustrates that even in this controlled setting most of the behaviors are decoded with 0.80 and higher AUCs. The best decoded behaviors remain to be EATING, VIDEO GAMES and SLEEP. The two worst decoded behaviors which have 0.7 AUCs are CONTACT for Subject 1 and PATIENT IS TALKING for Subject 3. Both these behaviors are very sparse and with short durations which might be the prime reason why they are not being decoded well.

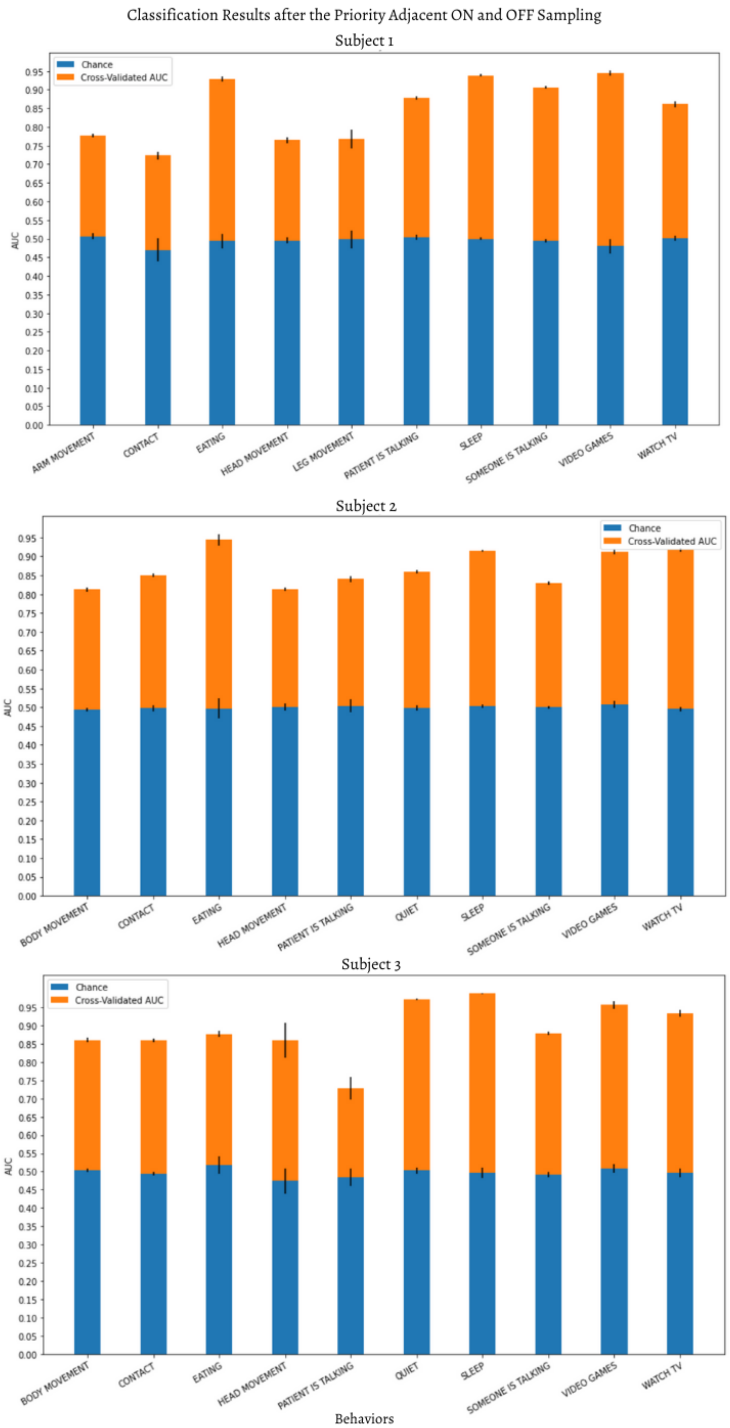

Figure 2.11: The plots showcase the classification AUCs after the data was sampled through a custom designed sampling method to prioritize the adjacent ON and OFF samples.

2.6.2 Time Invariance Analysis

In this section, I present the results of the analysis where the events were divided into two groups depending on whether they belonged to the first part of the day or the second. The events of the first part of the day are used for the training and the events of the second part of the day are used for testing. Please note that not all of the behaviors had enough events spread out during the day, so this test was possible only with the select number of behaviors which had good coverage during the day. The usual practices in Brain Computer Interfaces require training models in advance and use them to predict the outcomes for the future unseen events. For this reason, the models need to be tested for their robustness to handle the new events. Moreover, the annotations of this study are very general buckets for physiological behaviors. In fact, the behavior annotations might be too broad that the different events among these annotation buckets might not have homogeneous distributions over time. For example, BODY MOVEMENT includes both arm and leg movements which might have very distinct representations in the brain. If the training data has only leg movements and the test data has only arm movements, then the test classification quality may turn out much worse than expected. Or EATING breakfast might have a very different representation from EATING dinner.

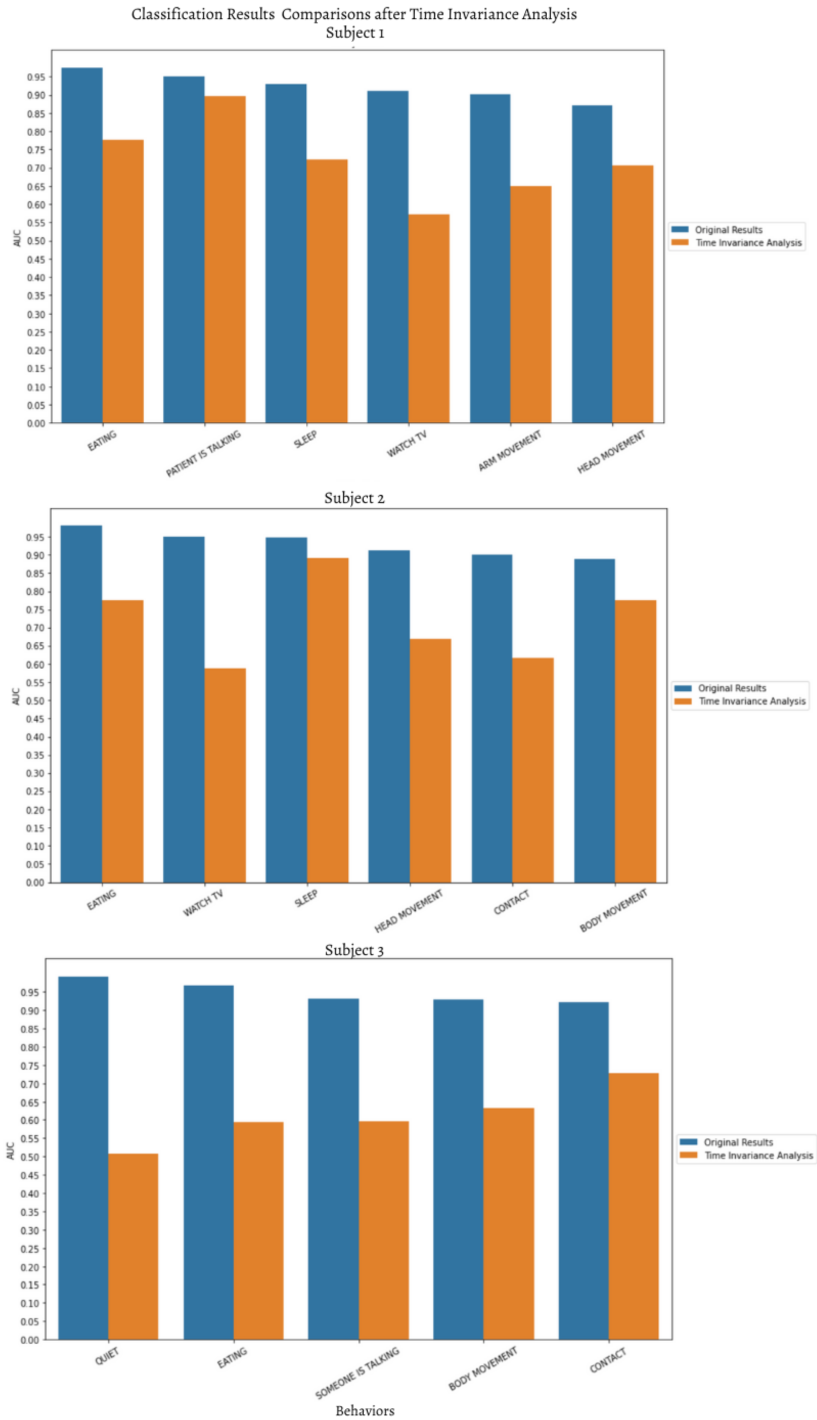

Figure 2.12: The plots show the comparisons of the original results for the behaviors and the classification results received for the time invariance analysis.

2.6.3 Sensitivity of Feature Engineering Design Choices

One major concern in the beginning of my research was to make sure the feature engineering is as precise and informative as possible. I wanted to make sure that I capture all the valuable information while also denoising the data as much as possible. As a result I came up with 4 different feature engineering methodologies where all of them had the valid justification to have sufficient decoding power. The first methodology, to which I refer as Methodology A in this paper, is where the windows were selected to be equally spaced and bipolar re-referencing was applied to the channels. Methodology B is similar to Methodology A, except I do not re-reference the channels. Both Methodology C and D adjust the windows to the start of the behavior changes where Methodology C also has bipolar re-referencing and Methodology D does not have re-referencing. In this section, I am presenting the classification AUC comparisons for the features calculated by the four aforementioned methodologies. As shown in Figure 2.13, the classification performance is not drastically different among the features, but Methodology A consistently outperforms the other methodologies. I expected that Methodology C and Methodology D will perform comparatively worse, as the window selection adjusted for the start of the behavior changes will increase the complexity of decoding, in the form of adding ambiguous and imperfectly annotated in-between behavior data points. Regarding the impact of re-referencing, I think bipolar re-referencing is marginally better than the original reference of the channels as Methodology A results are consistently better than those of Methodology B, and Methodology C is almost always better than Methodology D. Overall, it is reassuring to see that the different feature sets have comparable decoding performances, as it establishes once more that the classification results were not due to a lucky chance.

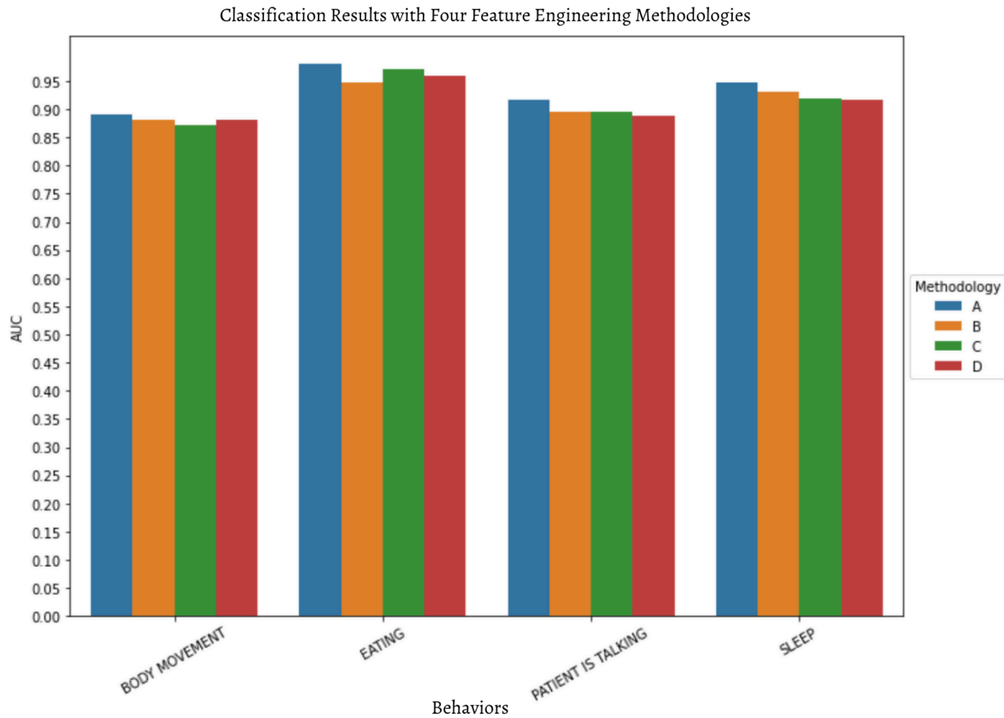

Figure 2.13: The plots show the comparisons of the classification results among the four engineered feature sets.

2.6.4 Window Selection Size

Selecting the windows to be roughly one second in the feature engineering process is another arbitrary choice which I made in my problem-solving process. Choosing one second as a length of the window was intuitive and easily interpretable, however, other reasonable choices for the window length were also available such as 2-second windows or half-second windows. If I chose two second windows, the data would be more denoised at the cost of losing precision. In this section, I test the sensitivity of my classification against a different window length selection. Figure 2.14 shows that in fact almost for all of the behaviors one second windows achieve better performance with their precision. However, the behaviors such as BODY MOVEMENT, HEAD MOVEMENT and PATIENT IS TALKING which are very noisy, with short event lengths and hard to annotate, the two second windows achieve better results. For these behaviors more denoising increases the classification AUCs.

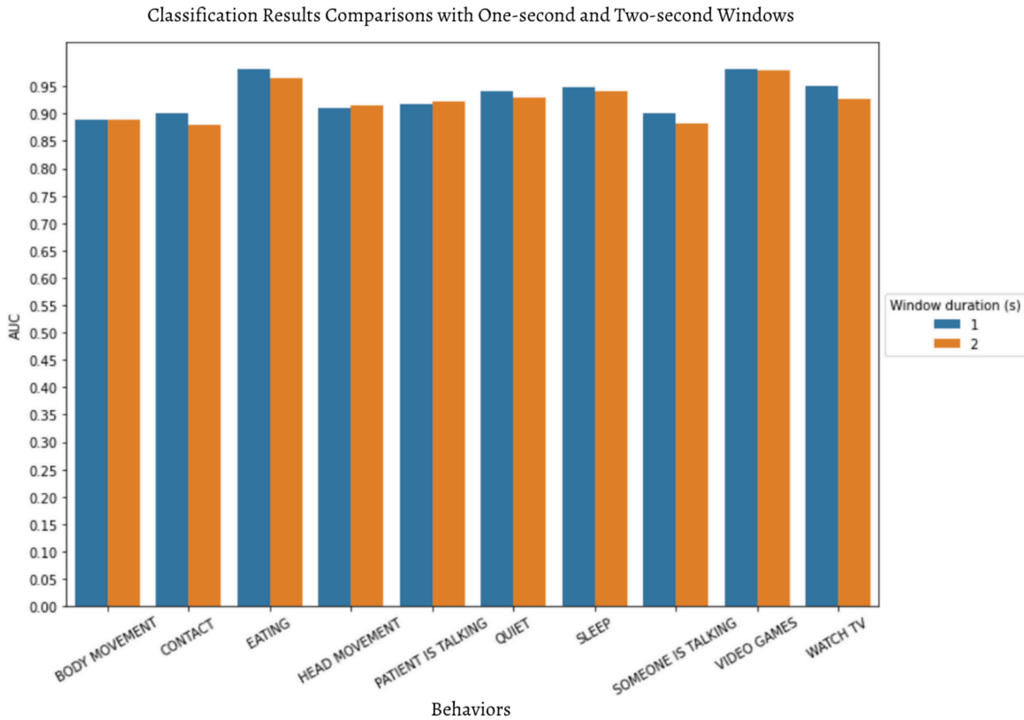

Figure 2.14: The plots show the comparisons of the classification results for the features calculated with one second windows vs features calculated with two second windows.

2.7 Feature Importance Analysis

Decoding human physiological behaviors also implies understanding the underlying driving factors which contribute to the classifier from the computational, feature engineering perspective, and from biological, brain region importance perspective. In this section I am demonstrating my feature importance analysis based on the Random Forest classifiers from the main results section. As discussed in Section 1.5.1 I use the Gini impurity score as my main feature importance metric, since the Random Forest models used that metric for the training optimization. Firstly, I wanted to know which features had the highest impact in decoding the distinct behaviors. Since I have calculated all of the nine features for each of the channels, I had to aggregate the Gini impurity scores across the channels for the corresponding feature. The Gini impurity scores are normalized across all the features such that the sum of all the importance scores over all the features is 1. Because of this normalization and the fact that all the features have equal number of corresponding channels to aggregate over, I simply summed the impurity scores for each feature across the channels. The results of this analysis are presented with heatplots in the Figures 2.15, 2.16, 2.17 where the y axis represents the behaviors and the x axis represents the features. Unfortunately, there is no consistent trend of feature importance across the subjects, as Subject 3 seems to rely entirely on RMS, MAV, MAX, MIN features; while Subject 1, on contrary, relies on mostly frequency-based features; and Subject 2 equally spreads the importance scores among time series and frequency-based features. These heatplots are also in sync with the time-series vs frequency-based classification results, where Subject 3 was the only patient for whom time-series based feature classifications were consistently outperforming frequency-based classification results for all the behaviors. However, I should also note that comparing the feature importance results for the subjects is hard, because each of the subjects has a different set/number of electrodes placed in different regions of the brain.

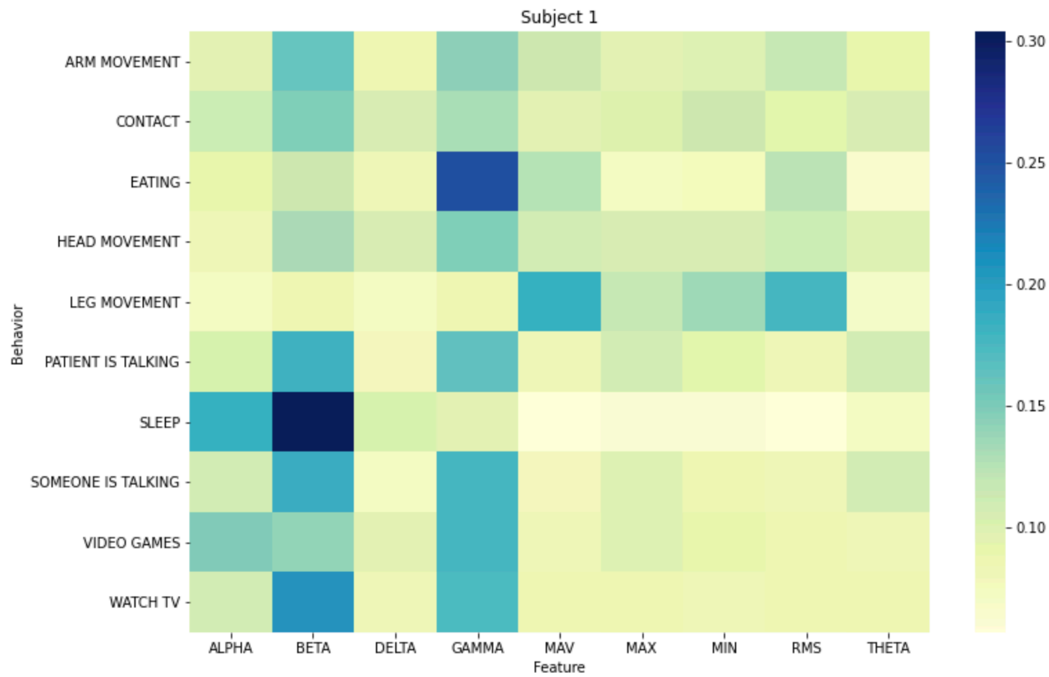

Figure 2.15: The heatmap in the figure illustrates the average importance of each feature in decoding the specific behaviors for Subject 1. The feature importance intensity is colorcoded and the scale is shown on the right.

Diving deeper into these results I have noticed some interesting patterns such as the fact that the GAMMA feature is very important for decoding EATING behavior for Subject 2 and 3, and also somewhat important for Subject 1. SLEEP behavior for Subject 1 is mainly decoded by ALPHA and BETA features, for Subject 2 the main importance is based on DELTA, MAX and MIN features; and Subject 3 seems to equally utilize GAMMA, MAX, MIN, MAV and RMS features. Generally, it seems that Subject 2 and Subject 3 share more similar feature importance score distribution than Subject 1. Even with WATCH TV behavior one can see that RMS has very high impact for Subject 2 and Subject 3, and almost no impact for Subject 1.

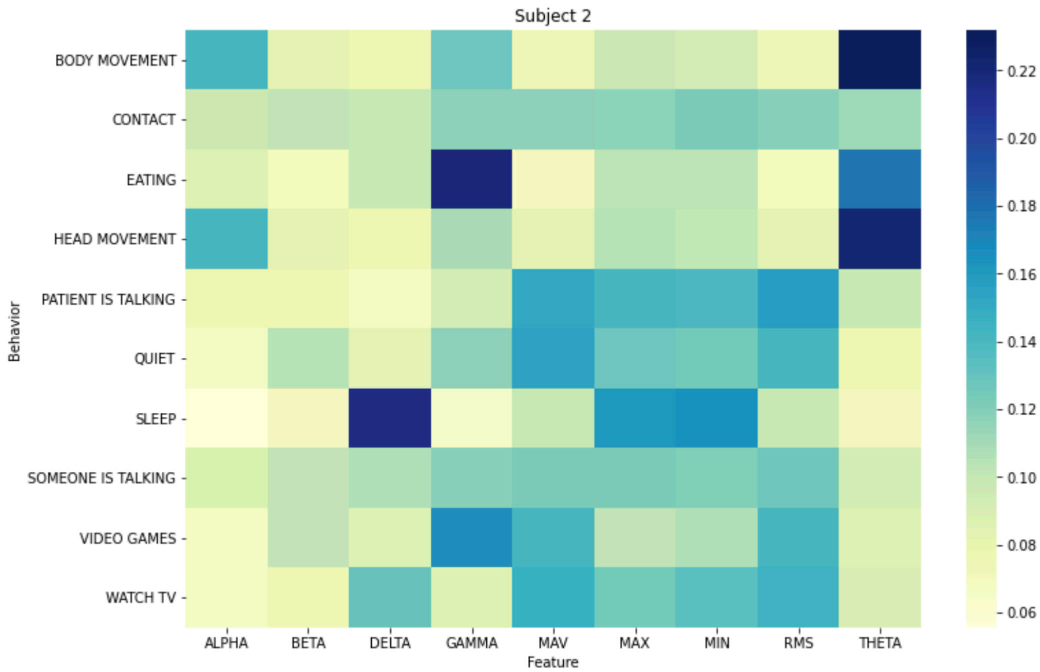

Figure 2.16: The heatmap in the figure illustrates the average importance of each feature in decoding the specific behaviors for Subject 2. The feature importance intensity is colorcoded and the scale is shown on the right.

Moving forward I wanted to scrutinize some of the behavior-feature pairs based on the brain regions where the channels were located. With this analysis I was mainly interested to see if the difference among the feature importances across the subjects was caused by the different number and placements of the initial electrodes. The first pair I had a look at was EATING-GAMMA, which was a strong pair for Subject 1 and Subject 2, and not that strong for Subject 3. The Figure 2.18 visualizes the brain regions on x axis labeled by their international abbreviations, the subjects on y axis and the feature importance score by the color intensity.

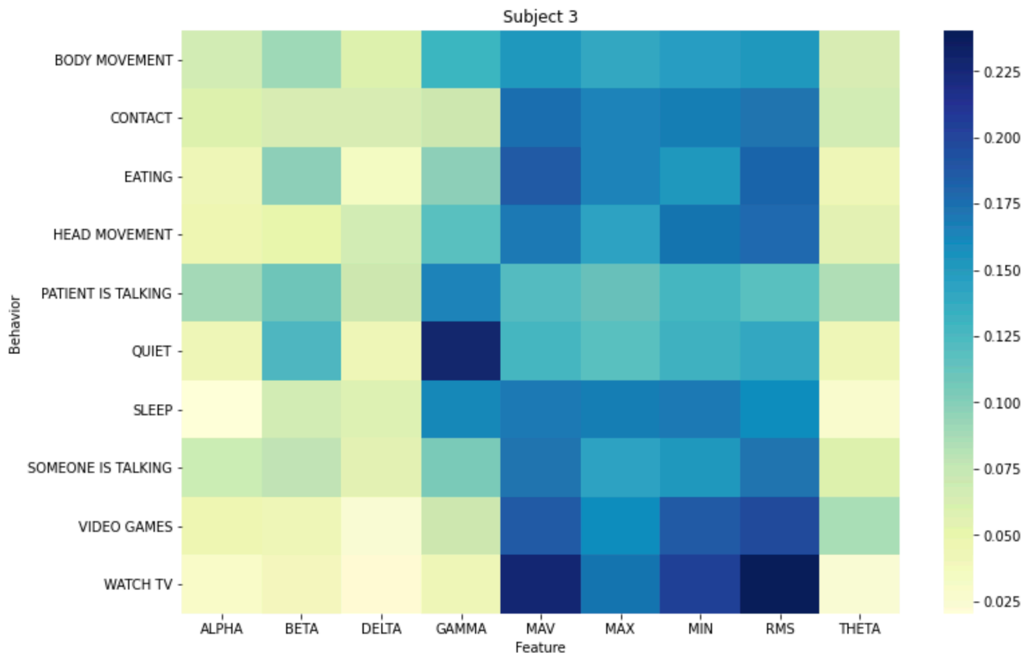

Figure 2.17: The heatmap in the figure illustrates the average importance of each feature in decoding the specific behaviors for Subject 3. The feature importance intensity is colorcoded and the scale is shown on the right.

Moreover, the white denotes the areas where the subject did not have electrode placements. Since for some of the regions there were a few channels, I aggregated the feature importance scores. This time I aggregated the scores by averaging because the number of channels for each brain region varied heavily from region to region and from subject to subject. From the first sight it becomes clear how much discrepancy there is in between the brain region placements for the subjects where Subject 3 has the most coverage(154 bipolar channels) and Subject one has the least coverage (35 bipolar channels). Looking further you will observe that the most important GAMMA features for Subject 2 were placed at the postcentral gyrus and both Subject 1 and Subject 3 miss electrodes in this brain region. The strongest GAMMA impact is observed from Subject 1 at the region of subcentral gyrus and sulcus. Subject 3 has no high GAMMA feature areas.

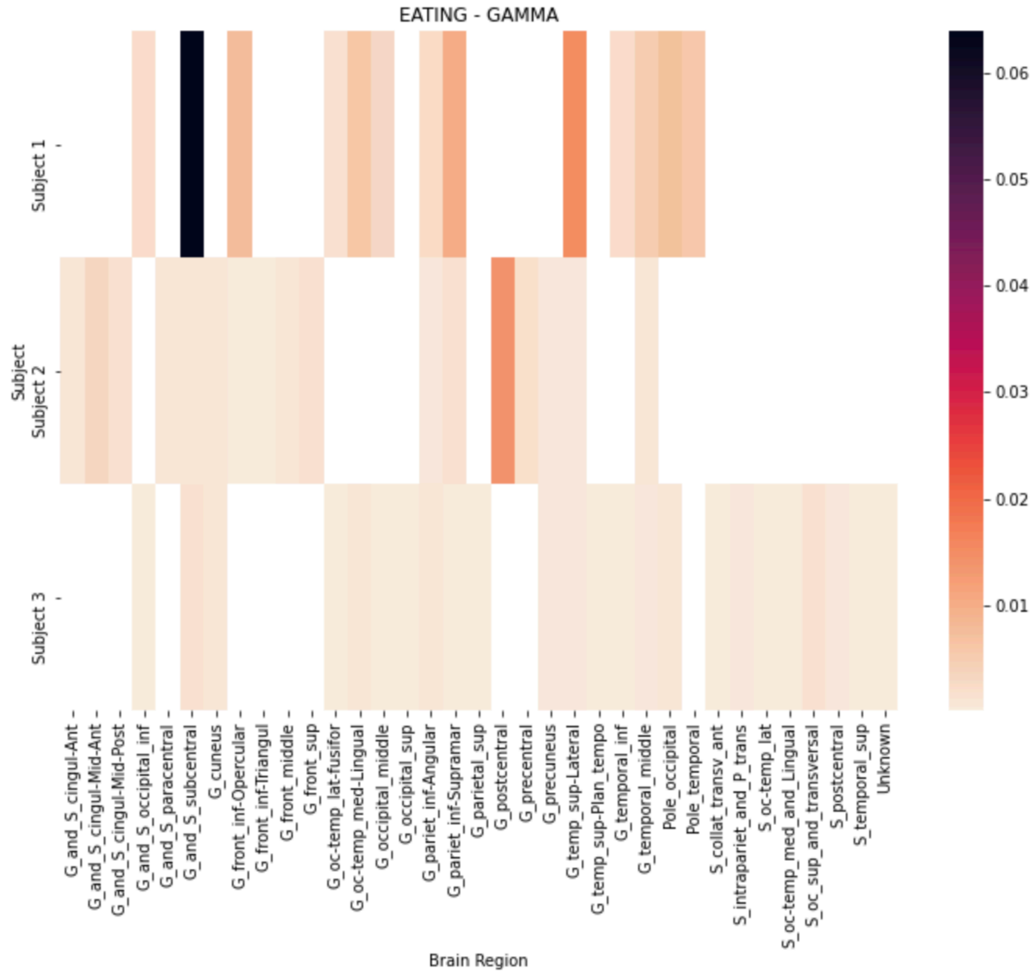

Figure 2.18: The heatmap in the figure illustrates the average importance of GAMMA features in decoding EATING behavior as it corresponds to brain regions. The feature importance intensity is colorcoded and the scale is shown on the right.

The next pair of analysis is the SLEEP-BETA pair as I was intrigued by how strong the BETA feature set was for decoding the SLEEP behavior for Subject 1 and how other subjects were not showing any feature importance for SLEEP-BETA. In fact Figure 2.19 one can see how the high importance BETA regions in Subject 1 exclusively do not exist for Subject 2. Subject 2 has almost no importance on BETA for decoding SLEEP and Subject 3 has very little. The high BETA importance areas for Subject 1 are occipital lobe sulci and gyri, fusiform gyrus or lateral occipitotemporal gyrus, lingual gyrus or the medial occipitotemporal gyrus, middle occipital gyrus and occipital pole.

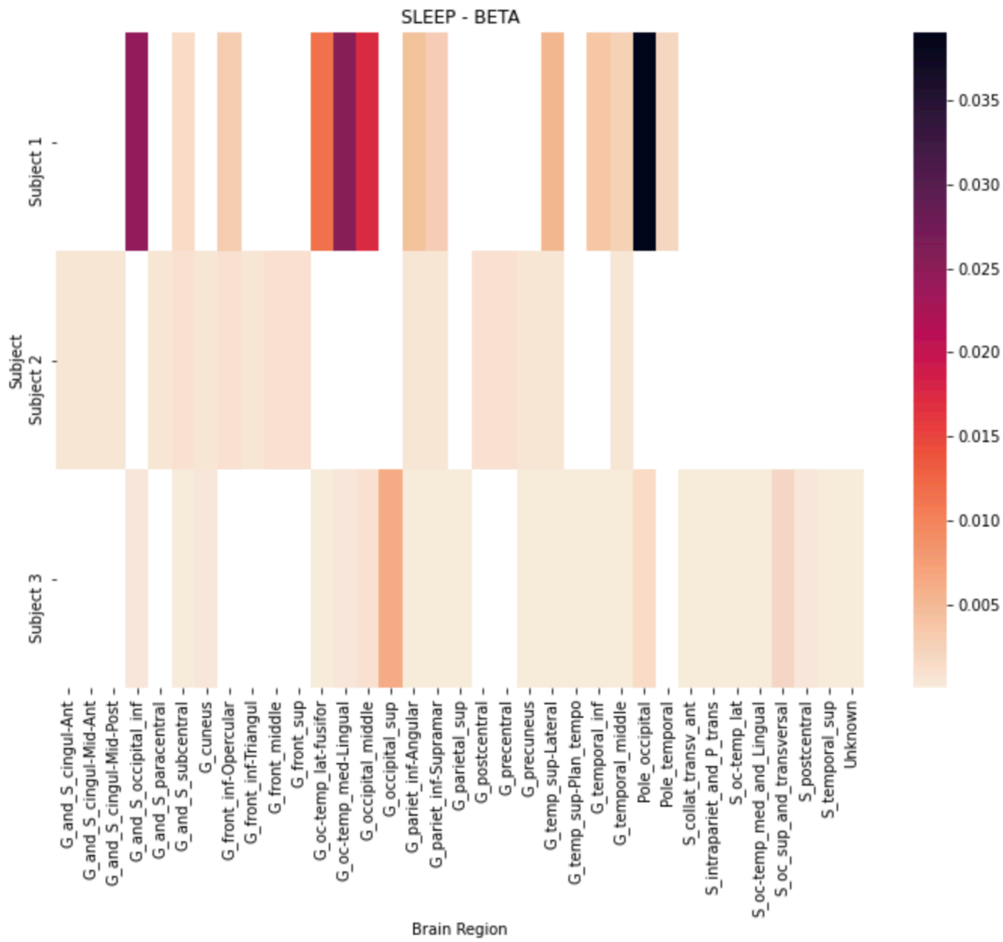

Figure 2.19: The heatmap in the figure illustrates the average importance of BETA features in decoding SLEEP behavior as it corresponds to brain regions. The feature importance intensity is colorcoded and the scale is shown on the right.

I have also investigated the WATCH TV - RMS pair, as in the overall behavior-feature importance score heatmaps the RMS features had high importance scores for Subject 2 and Subject 3, but not for Subject 1. The highest importance score occurs for Subject 3 at the precuneus gyrus and that area had no electrode placements for Subject 1. Having said that even though Subject 2 had electrode placements in that area, the feature importance score from that area is very low for RMS features. Moreover, all the other high impact RMS areas for Subject 3, such as cuneus gyrus, transverse occipital sulcus and postcentral sulcus, also miss from Subject 1's and partially from Subject 2's electrode placements. This shows that the difference in electrode placements makes it really difficult to compare the feature importances among the patients.

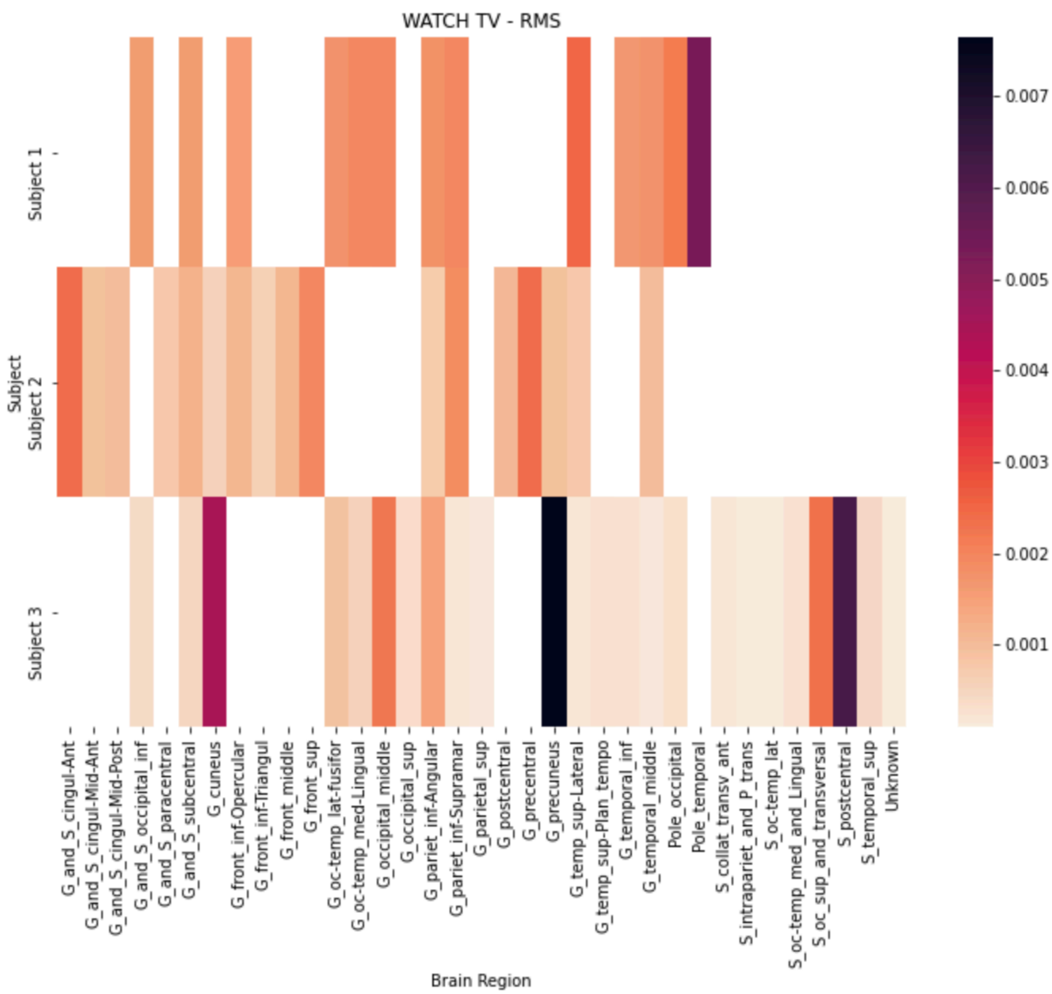

Figure 2.20: The heatmap in the figure illustrates the average importance of RMS features in decoding WATCH TV behavior as it corresponds to brain regions. The feature importance intensity is colorcoded and the scale is shown on the right.

Finally, I was curious to see how much each brain region contributed to the classification success for each of the behaviors, so I visualized Figures 2.21-23 for all the subjects where y axis represents the behaviors, x axis represents the brain region and the color intensity represents the average feature importance scores. When interpreting the figures vertically, the occipital gyrus and sulcus, superior temporal gyrus and occipital pole have on average higher impact for most of the behavior for Subject 1. The region behavior pairs that stand out the most from Subject 1 are the subcentral gyrus and sulcus impact on the EATING behavior; the lingual gyrus and VIDEO GAMES; the middle occipital gyrus and LEG MOVEMENTS; occipital pole and WATCH TV.

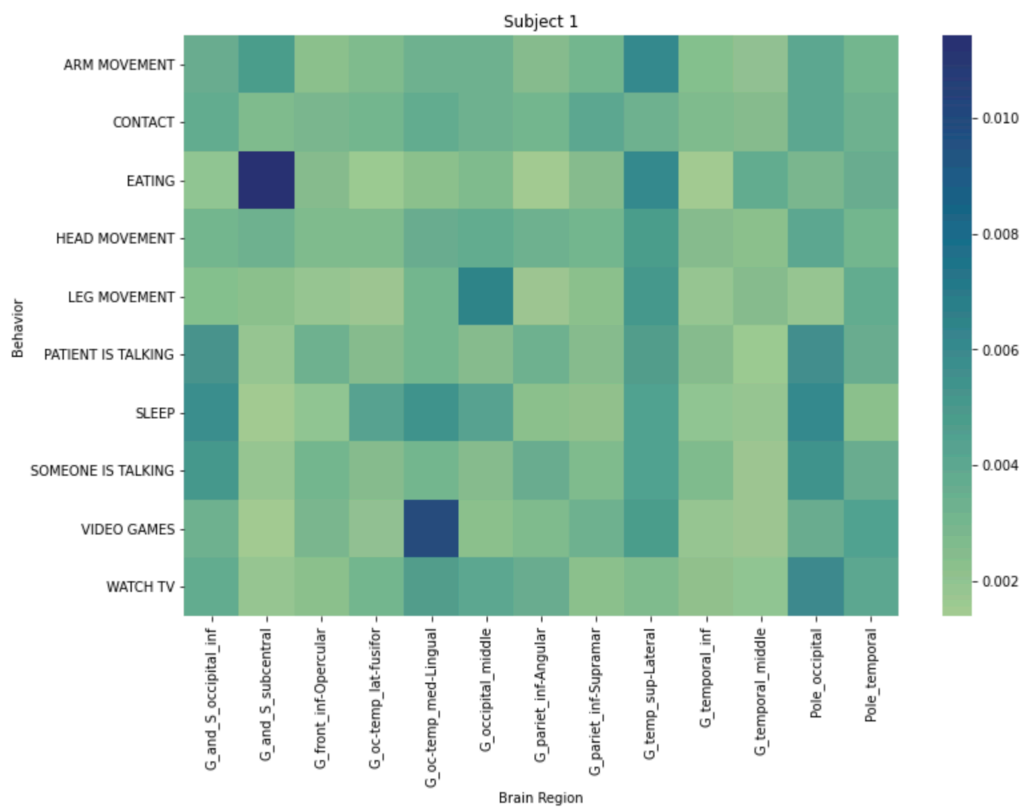

Figure 2.21: The heatmap in the figure illustrates the averaged feature importance scores by brain regions for Subject 1. The feature importance intensity is colorcoded and the scale is shown on the right.

For Subject 2 each behavior seems to have brain regions that clearly contribute more than the other regions. Paracentral gyrus and sulcus and postcentral gyrus contribute the most among the brain regions to BODY MOVEMENT decoding. Anterior midcingulate gyrus and sulcus, subcentral gyrus and sulcus, postcentral gyrus and temporal middle gyrus are all high feature importance areas for EATING behavior. According to the figure, HEAD MOVEMENT has high importance scores at the region of paracentral gyrus and sulcus and postcentral gyrus. Postcentral, precentral and subcentral gyrus are active in decoding SLEEP. Lastly, the frontal superior gyrus has high impact for VIDEO GAMES behavior in Subject 2 classifications.

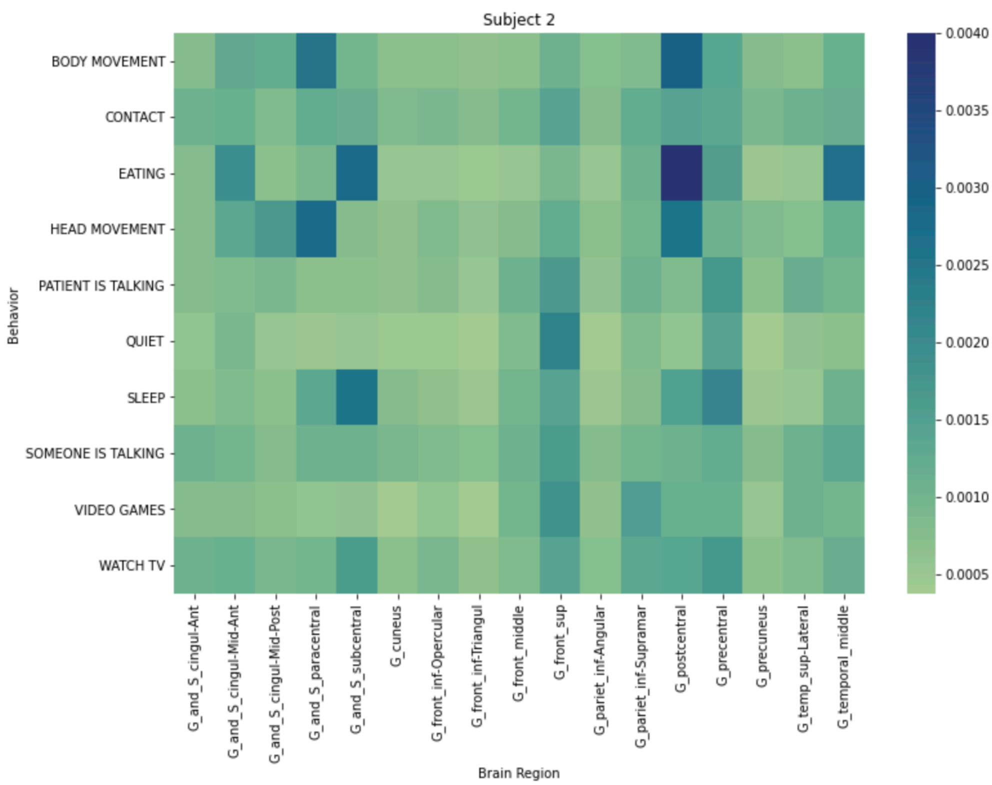

Figure 2.22: The heatmap in the figure illustrates the averaged feature importance scores by brain regions for Subject 2. The feature importance intensity is colorcoded and the scale is shown on the right.

Since our third patient had the largest number of implanted electrodes the Figure 2.23 has more brain region coverage. Cuneus and precuneus giri in Subject 3 analysis seem to have higher than average impact for all the behavior classifications. Cuneus and precuneus giri were also the regions where 20% of the electrode implants were placed, so one can imagine that there was a lot of information obtained from these regions. The CONTACT behavior seems to be decoded well from the information received from subcentral gyrus and sulcus. The superior occipital gyrus is the region with the highest feature importance score for HEAD MOVEMENT. The postcentral sulcus with middle occipital gyrus have high impact on QUIET and SLEEP behaviors. Lastly, WATCH TV has high importance scores for precuneus gyrus and postcentral sulcus.

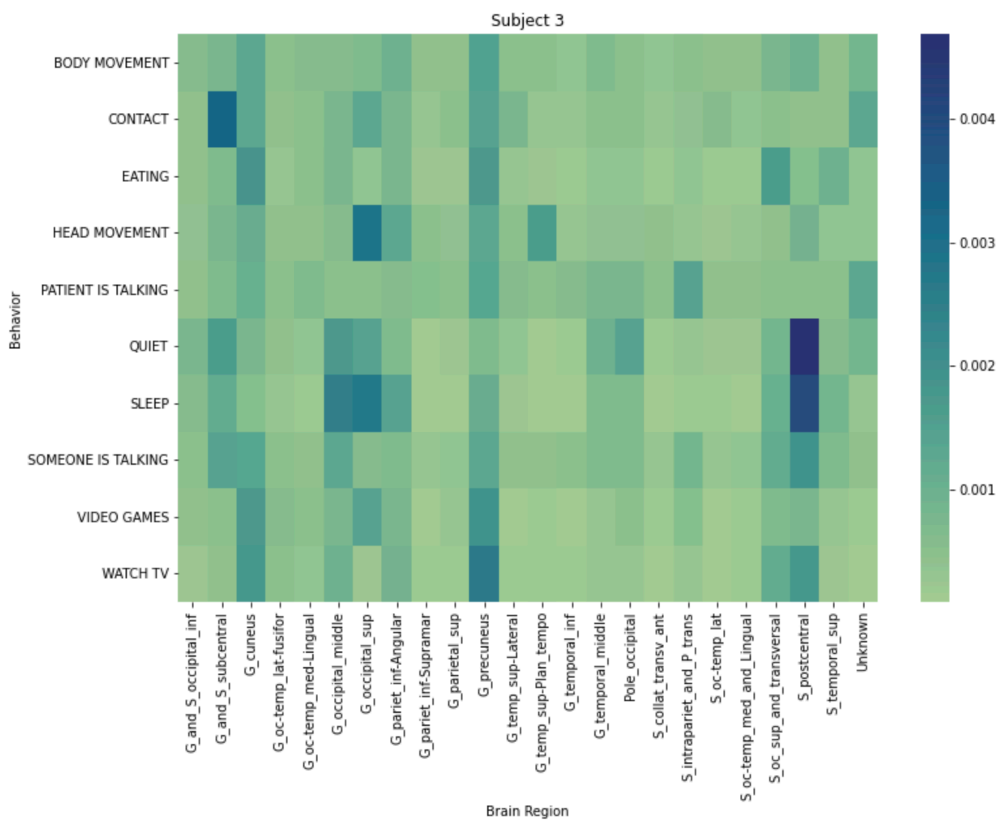

Figure 2.23: The heatmap in the figure illustrates the averaged feature importance scores by brain regions for Subject 3. The feature importance intensity is colorcoded and the scale is shown on the right.

When I look at the human brain I'm still in awe of it.

Ben Carson

3
Discussion

3.1 Main takeaways

Throughout the results section I have presented the decoding performance results for the various behaviors across all the three subjects, and most importantly, my results showed that it is possible to dynamically classify the continuous human physiological behaviors over time. The annotated behaviors can be divided into two groups: behaviors that are characterized with longer lasting events such as SLEEP, QUIET, SOMEONE IS TALKING and WATCH TV; and short but more frequent events such as the movement based behaviors of BODY MOVEMENT, LEG MOVEMENT, ARM MOVEMENT, HEAD MOVEMENT. Generally, the longer lasting events get decoded better since there is less noise in both annotation precisions and from the brain activity perspective these states are better defined since they last longer for our sampling frequency to capture their unique features. Some examples of these good performances are 0.90 and higher AUCs obtained while decoding SOMEONE IS TALKING and WATCH TV behaviors, 0.92 and higher AUCs for decoding SLEEP, 0.93 and higher AUCs for decoding QUIET. The movement based behaviors are usually around 0.9 AUCs even though for some cases such as BODY MOVEMENT for Subject 3 is decoded with 0.93 AUC.

The two outliers that do not comply with the aforementioned discussed pattern are EATING and VIDEO GAMES behaviors. From the figure which shows the ON and OFF annotations for behaviors over the time (Figure 2.1), one can see that neither EATING, nor VIDEO GAMES happen often or last longer, and yet, they are decoded with the highest AUCs (0.97 and 0.95 respectively). Both behaviors require engaging multiple sensory receptors simultaneously creating rich sensory stimuli for the brain and a component of consciously acting on the input information. From the prior literature it is shown that the sensory stimuli, in fact, are linked and have high correlations with brain activity at certain brain regions which are responsible for the successful processing of this input information.[18] EATING kindles visual, olfactory and gustatory systems to work all together, while VIDEO GAMES requires the patient to process the audio-visual input and put a conscious mental effort on deciding what to do in the game according to the input information. All these considerations might have created distinct neural representations in the brain which helped with decoding these activities with higher AUCs. In the series of experiments I tried to test the sensitivity of my models introducing various controls to which my models decreased their AUCs but still remained significant showing that the decoding success was not due to a lucky chance. After implementing the adjacent ON and OFF sampling discussed in Section 1.8.2, the classification AUCs decreased on average by 7% across the behaviors, but this average is a little misleading since this decrease is mainly due to the drastic decrease in short/more frequent behaviors such as the movements. SLEEP, SOMEONE IS TALKING, EATING and VIDEO GAMES were more robust to the adjacent sampling and had decreased by at most 3%. These results were not at all surprising, because since the beginning I knew that movement based activities were noisy, and prioritizing the samples which were right in between ON and OFF events would introduce even more noise in the classification process, worsening the performance.

The sensitivity tests on comparing the classification performances using different window sizes (one second vs two second windows) and the different feature sets were robust across all the behaviors and subjects ensuring once again that the results of this research capture a real effect. It is noteworthy to mention that increasing the window size implied losing precision but also had the effect of additional denoising, which is consistent with the results of this sensitivity test (Figure 2.14) where the frequent/noisy behaviors were the only behaviors to see increase in their classification performances for the features computed with two second windows.

Lastly, since my motivations for this research were to get as close as possible to possible real-life deployments of these classifiers I have introduced a test which mimicked a scenario similar to BCI model testing process (Section 1.8.2). In this experiment I test if my model can robustly classify the future events which can be out of distribution for my training pool (OOD), if the training data comes only from one part of the day. The results of this test are humbling as there is a huge drop in classification AUCs (by 2.6.3). I was expecting to see a decrease in performance as since the beginning of this research one of the concerns was that the chosen buckets for behavior annotations were not specific enough, and hence, might have resulted in an inhomogeneous distribution of events. For example, BODY MOVEMENT behavior itself is defined to be very broad - it can be leg movement, arm movement. And even for ARM MOVEMENT and LEG MOVEMENT there might be a differentiation in our brain motor system for the right and left movements which are not taken into account. Moreover, for the behaviors which engage with sensory information such as EATING, the EATING in the morning might be a different sensory experience from the EATING in the evening. All of these considerations are examples of out of distribution problems in machine learning. Furthermore, not having enough of these events during our day will hinder the classifier from learning the high level general patterns for EATING and fail in the testing process. Despite all the concerns and the decrease in the performance AUCs, I should state that the classification performances are still significant meaning that the classifiers are still able to differentiate among the ON and OFF behaviors significantly better than the random model. The general pattern is that the behaviors which were frequent and had many events throughout the day get decoded better under this experiment constraint. However, an interesting outlier in this test is the EATING behavior for which the classifier not only decreases the accuracy but performs even worse than the random model. The reason behind this anomaly might be that the EATING behavior events in the train set have a different representation in the brain compared to the EATING events in the test set and the classifier fails to predict those out of distribution events. Gathering more data and tracking the subjects over longer periods of times to get more EATING events will alleviate this problem.

3.2 Feature Importance

While showing that different physiological behaviors can be successfully decoded is one of the main aims of this paper, it also strives to understand the how and why behind the achieved results. From the computational point of view it is important to validate which features were the most effective in representing useful information from the noisy voltage waves that we started from. For example, if there was a consistent pattern in the feature importance scores for certain groups of features (e.g.

RMS, DELTA), I could recommend the future researchers to calculate those features in their feature engineering process. However, as the feature importance heatmaps showed in Section 2.7 there is no clear pattern across the subjects or behaviors about dominantly important feature sets that could be a point of recommendation. Given that all of my subjects had different numbers and locations of electrode placements, the comparisons of the feature importances are not very trustworthy across the subjects. Observing the subjects and behaviors on a case by case basis, I found that for Subject 1 the more important features across the behaviors were ALPHA, BETA and GAMMA. Subject 2 had a mix of important features depending on the decoded behavior. And finally, for Subject 3 the time-series based features were the most effective.

3.3 Brain Regions

From the prior literature it is known that different regions in the human brain correspond to different biological functions. Even though the implanted electrodes did not cover all the regions of the cerebral cortex for all the patients, there is still good coverage of the cortex which can be used for exploring which brain regions contributed to the decoding success of the different physiological behaviors. As an overview, the majority of the electrodes for Subject 1 were implanted in the temporal lobe, which is linked with speech, short-term memory and smell recognition. Most of Subject 2's electrodes were placed in the frontal lobe, which is mainly associated with decision making and movement. And finally, the vast majority of Subject 3's electrodes were implanted in the occipital lobe, the part of the brain that is responsible for vision.

When analyzing the feature importance scores based on the brain regions I found that to decode EATING behavior for Subject 1 the most importance was put on the Subcentral Gyrus which in various researches was linked to speech-related mouth movements.[16] While EATING behavior in our annotations does not imply any kind of speech, it definitely implies mouth movements for food consumption, so the heavy importance of this region can be somewhat justified and linked to the previous research studies. For Subject 2 the most important features in decoding EATING were from Postcentral and Subcentral Gyrus. Postcentral Gyrus is known to be the main brain region responsible for the sense of touch, which explains why EATING behavior being one of the richest sensory experiences would benefit from information provided by this region.[19] EATING for Subject 3 equally placed importances on Cuneus Gyrus, Precuneus Gyrus and Superior Occipital Gyrus. Cuneus Gyrus is known for its basic visual processing functionalities.[9] Precuneus Gyrus is involved in complex functions such as recollection of memories and integration of information relating to the perception of the environment.[21] Superior Occipital Gyrus is mainly connected with object recognition but recent research has shown that this gyrus also plays roles in reaching and grasping an object of desire.[17] Since EATING implies that there were food objects which could serve as visual stimuli, for the classifier to place importance on the gyri responsible for visual information processing is justified, especially, given the fact that most of the electrodes for Subject 3 were placed in the occipital cortex and there was not that much choice of diversity.

The VIDEO GAMES behavior in Subject 1 put a lot of emphasis on Lingual gyrus, which is a region associated with the global shape processing of the visual stimuli, especially when it is a word or letter.[28] Since gaming is all about receiving visual stimuli and responding to it, these were not surprising. For Subject 2 VIDEO GAMES were best decoded by using the information from Superior Frontal Gyrus which is thought to be a major contributor to working memory, a functionality necessarily used in gaming.[17] Lastly, for Subject 3 Cuneus, Precuneus and Superior Occipital Gyri had equally average importance scores. As mentioned before all of these regions are in one way or another linked to object detection and visual information processing which are again intuitively important in VIDEO GAMES behavior.

While SLEEP is not a behavior that is localized in any of the brain regions, its characteristics such as having eyes closed, no visual inputs and minimal movement are clues that can be used by the classifier while utilizing the corresponding motor and visual information processing regions of the brain. Additionally, for the SLEEP I looked at the frequency based features since it was shown that the SLEEP is positively associated with DELTA frequencies and, hence, negatively correlated with all the frequency bands that are associated with being awake such as BETA or GAMMA.[5] Among all three subjects the main feature importance areas were Inferior, Middle, Superior Occipital Gyri, Postcentral gyrus and Subcentral gyrus. Additionally, when looking at SLEEP-BETA pairs, Subject 1 showcases very high BETA importance in Pole Occipital. Occipital complex, as mentioned before, was responsible for object detection, so with the inverse logic, no object being detected could be successfully associated with the SLEEP state by the classifier. Lastly, among the classifiers of the movement related behaviors only Subject 2 has highly differentiated brain region areas of importance for BODY MOVEMENT. Specifically, Paracentral Gyrus and Postcentral Gyrus are where the highest feature importance scores are located on average. As mentioned earlier, the Postcentral Gyrus is the main area for the sense of touch, which will be an active component for the type of movements the patient does, since they lie in the bed and most of the movements are limited: they either reach for something or move in the bed. Paracentral Gyrus is the continuation of the Postcentral and Precentral gyri and is responsible for important motor innervations as it relates to lower limbs and perineum.[32] In fact, Subject 2 is the only subject that had electrodes implanted in this area, which makes sense why this brain region's importance was emphasized only for Subject 2.

3.4 Sources of Error in Classifications

I mentioned multiple times in the paper that the ECoG recordings used in decoding the physiological behaviors were very noisy, and potentially, caused a lot of uncertainty in the classification process. Additionally, the annotations which were paired with the iEEG data and used for supervised learning came with a lot of errors. The most obvious error could be due to human reaction time, as I did not expect the annotators to have a perfect millisecond precision. In fact, the feature engineering design for aggregating the neuronal recordings in one second windows and getting the main characteristics of the voltage wave in that window was a design choice to alleviate the aforementioned error. The annotations also suffered from imprecisions such that the ON event for a behavior would be followed with another ON event without having an OFF interruption in between. Even though these human errors were not a substantial part of the annotations, they still could be a potential source of error in the classification. Another problematic factor in the classifications was the choice of annotated behaviors and how homogenous the different ON events were for those behaviors. As discussed before, BODY MOVEMENT includes many types of movements which are represented differently in the brain and create inhomogeneous event distributions for BODY MOVEMENT. With this problem the classifier might not be able to generalize well on the events of BODY MOVEMENT which were not present in the training set. When thinking of the ECoG data, something that came up a lot in the discussion of the results was the fact that the patients had electrodes implanted only in one part of their brain so it was hard to compare the results across the patients and draw conclusions. In general there was not enough coverage of the brain regions for all of the subjects which could also serve as a source of error in classification of the behaviors. Finally, even though my data was very generous spanning 24 hours there were behaviors which had only a few occurrences during the day such as EATING. To have a better confidence on the model and the results, it would be more preferable to track the subject for a longer period of time and collect more events of the underrepresented behaviors.

3.5 Time and Efficiency

This study has been very computationally costly as the intracranial field potentials data combined for the three patients occupied more than 150 GB memory to start with. Despite all the parallelization efforts the feature engineering took a significant amount of time: the computation time for calculating a single feature for a single channel took on average 20 minutes. For the comprehensiveness of this study, I have designed and calculated 9 sets of features (RMS, MAV, MIN, MAX, ALPHA, BETA, GAMMA, DELTA, THETA) with 4 different Methodologies. Each feature set was calculated for 35, 91 and 154 channels corresponding to Subject 1, Subject 2 and Subject 3 respectively with bipolar re-referencing, and 40, 104 and 176 channels without the re-referencing. The feature engineering designs of aggregating of 250 samples per one window was also aimed at making the decoding computationally more efficient. Even though the dataset was largely reduced after the feature computations, it was still computationally heavy for Random Forests to fit 800 estimators with depth 20, so the model training took a sizable amount of time as well.

3.6 Data Ethics

Since the study involves data directly received from human subjects it is important to note that all of the subjects have been anonymized and no personally identifiable information(PII) was used in the study which could put the patients' privacy under a risk. Moreover, the data was approved for the research study by the Institutional Review Boards at Boston Children's Hospital and the patients consented to be video recorded. The main motivation of this study has been and remains to take a step forward in BCI research for its medical and health-related applications.

4

Conclusion and Future Work

This study establishes that the decoding of various continuous physiological behaviors from the information received directly from the brain is possible in uncontrolled environments. It also confirms the consistency of representations of these physiological processes in their corresponding brain regions as supported in prior literature. The well-defined behaviors which lasted for longer periods of time were generally classified with higher performance than the behaviors with quicker and more frequent events. All of the physiological behaviors among the three subjects have demonstrated consistently robust decoding performance for the various sensitivity tests discussed in Section 2.6. These findings suggest that the success of the classifiers is not merely due to chance but through extracting and analyzing relevant information from the brain.

To develop the study further and increase the confidence of the results the data preprocessing will need to establish a more rigorous annotating process. This can be achieved by having more annotators for the same video-recordings or by deploying a combination of human-AI annotating system. New experiments working on similar decoding tasks should include more coverage of the brain to get a better understanding of which brain regions are relevant in decoding the certain behaviors. Moreover, the study should be conducted with more subjects and for longer periods of time to make sure the discovered patterns are generalizable and representative of a larger population. Lastly, both data acquisition and analysis are very computationally costly for this experiment, so optimizing the usage of computational resources is vital for advancing the state of the art models in these types of research.There is still a long way to go until this study could be fully applied in BCI solutions but its findings hold a promise of a path for more robust brain decoding.

References

[1] (2021). Gyri and sulci of the brain.

[2] (2022). Brain anatomy and how the brain works.

[3] (2022). Cerebral cortex.

[4] Abhang, P. A., Gawali, B., & Mehrotra, S. C. (2016a). *Introduction to EEG-and speech-based emotion recognition*. Academic Press.

[5] Abhang, P. A., Gawali, B. W., & Mehrotra, S. C. (2016b). Technical aspects of brain rhythms and speech parameters. *Introduction to EEG-and speech-based emotion recognition*, (pp. 51–79).

[6] Aminoff, M. J., Boller, F., Bruyn, G., Klawans, H. L., Swaab, D. F., & Vinken, P. (1968). *Handbook of clinical neurology*. North-Holland Publishing Company.

[7] Bewick, V., Cheek, L., & Ball, J. (2004). Statistics review 13: receiver operating characteristic curves. *Critical care*, 8(6), 1–5.

[8] Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5–32.

[9] Bullard, S. E., Griss, M., Greene, S., & Gekker, A. (2013). Encyclopedia of clinical neuropsychology.

[10] Buzsáki, G., Anastassiou, C. A., & Koch, C. (2012). The origin of extracellular fields and currents—eeg, ecog, lfp and spikes. *Nature reviews neuroscience*, 13(6), 407–420.

[11] Clancy, E. & Hogan, N. (1997). Theoretic and experimental comparison of root-mean-square and mean-absolute-value electromyogram amplitude detectors. In *Proceedings of the 19th Annual International Conference of the IEEE Engineering in Medicine and Biology Society.'Magnificent Milestones and Emerging Opportunities in Medical Engineering'(Cat. No. 97CH36136)*, volume 3 (pp. 1267–1270).: IEEE.

[12] Council, N. R. et al. (2008). Emerging cognitive neuroscience and related technologies.

[13] Czeisler, C., Buxton, O., Kryger, M., Roth, T., & Dement, W. (2015). Principles and practice of sleep medicine sixth edition.

[14] de Cheveigné, A. & Nelken, I. (2019). Filters: when, why, and how (not) to use them. *Neuron*, 102(2), 280–293.

[15] Derix, J., Iljina, O., Weiske, J., Schulze-Bonhage, A., Aertsen, A., & Ball, T. (2014). From speech to thought: the neuronal basis of cognitive units in non-experimental, real-life communication investigated using ecog. *Frontiers in human neuroscience*, 8, 383.

[16] Eichert, N., Papp, D., Mars, R. B., & Watkins, K. E. (2020). Mapping human laryngeal motor cortex during vocalization. *Cerebral Cortex*, 30(12), 6254–6269.

[17] Galletti, C., Kutz, D. F., Gamberini, M., Breveglieri, R., & Fattori, P. (2003). Role of the medial parieto-occipital cortex in the control of reaching and grasping movements. *Experimental brain research*, 153(2), 158–170.

[18] Goldberg, R. F., Perfetti, C. A., & Schneider, W. (2006). Perceptual knowledge retrieval activates sensory brain regions. *Journal of Neuroscience*, 26(18), 4917–4921.

[19] Johns, P. (2014). Chapter 3 - functional neuroanatomy. In P. Johns (Ed.), *Clinical Neuroscience* (pp. 27–47). Churchill Livingstone.

[20] Jolliffe, I. T. & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065), 20150202.

[21] Josipovic, Z. (2019). Nondual awareness: consciousness-as-such as non-representational reflexivity. *Progress in brain research*, 244, 273–298.

[22] Klimesch, W. (2012). Alpha-band oscillations, attention, and controlled access to stored information. *Trends in cognitive sciences*, 16(12), 606–617.

[23] Kropotov, J. (2016). *Functional neuromarkers for psychiatry: Applications for diagnosis and treatment*. Academic Press.

[24] Lee, J., Leung, V., Lee, A.-H., Huang, J., Asbeck, P., Mercier, P. P., Shellhammer, S., Larson, L., Laiwalla, F., & Nurmikko, A. (2021). Neural recording and stimulation using wireless networks of microimplants. *Nature Electronics*, 4(8), 604–614.

[25] Lei, X. & Liao, K. (2017). Understanding the influences of eeg reference: a large-scale brain network perspective. *Frontiers in neuroscience*, 11, 205.

[26] Leuthardt, E. C., Schalk, G., Moran, D., & Ojemann, J. G. (2006). The emerging world of motor neuroprosthetics: a neurosurgical perspective. *NEUROSURGERY-BALTIMORE THEN HAGERSTOWN MD-*, 59(1), 1.

[27] Lotte, F., Nam, C. S., & Nijholt, A. (2018). Introduction: evolution of brain-computer interfaces.

[28] Mechelli, A., Humphreys, G. W., Mayall, K., Olson, A., & Price, C. J. (2000). Differential effects of word length and visual contrast in the fusiform and lingual gyri during. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 267(1455), 1909–1913.

[29] Moini, J. & Piran, P. (2020). *Functional and Clinical Neuroanatomy*. Elsevier.

[30] Nayak, C. S. & Anilkumar, A. C. (2019). Eeg normal waveforms.

[Norcliffe-Kaufmann et al.] Norcliffe-Kaufmann, L., Axelrod, F. B., & Gutierrez, J. Encyclopedia of neurological sciences 2nd edition.

[32] Patra, A., Kaur, H., Chaudhary, P., Asghar, A., & Singal, A. (2021). Morphology and morphometry of human paracentral lobule: an anatomical study with its application in neurosurgery. *Asian Journal of Neurosurgery*, 16(2), 349.

[33] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

[34] Ramachandran, V. S. (2012). *Encyclopedia of human behavior*. Academic Press.

[35] Rasheed, S. (2021). A review of the role of machine learning techniques towards brain–computer interface applications. *Machine Learning and Knowledge Extraction*, 3(4), 835–862.

[36] Saeidi, M., Karwowski, W., Farahani, F. V., Fiok, K., Taiar, R., Hancock, P., & Al-Juaid, A. (2021). Neural decoding of eeg signals with machine learning: A systematic review. *Brain Sciences*, 11(11), 1525.

[37] Sirven, J. (2014). Electrocorticogram (ecog). In M. J. Aminoff & R. B. Daroff (Eds.), *Encyclopedia of the Neurological Sciences (Second Edition)* (pp. 1080–1083). Oxford: Academic Press, second edition edition.

[38] Sörnmo, L. & Laguna, P. (2005). *Bioelectrical signal processing in cardiac and neurological applications*, volume 8. Academic Press.

[39] Tang, H., Yu, H.-Y., Chou, C.-C., Crone, N. E., Madsen, J. R., Anderson, W. S., & Kreiman, G. (2016). Cascade of neural processing orchestrates cognitive control in human frontal cortex. *Elife*, 5, e12352.

[40] Tudor, M., Tudor, L., & Tudor, K. I. (2005). Hans berger (1873-1941)–the history of electroencephalography. *Acta medica Croatica: casopis Hravatske akademije medicinskih znanosti*, 59(4), 307–313.

[41] Van der Maaten, L. & Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

[42] Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., & SciPy 1.0 Contributors (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17, 261–272.

[43] Wang, N. X., Olson, J. D., Ojemann, J. G., Rao, R. P., & Brunton, B. W. (2016). Unsupervised decoding of long-term, naturalistic human neural recordings with automated video and audio annotations. *Frontiers in human neuroscience*, 10, 165.

[44] Wang, Z., Gunduz, A., Brunner, P., Ritaccio, A. L., Ji, Q., & Schalk, G. (2012). Decoding onset and direction of movements using electrocorticographic (ecog) signals in humans. *Frontiers in neuroengineering*, 5, 15.

[45] Wolpaw, J. R. & McFarland, D. J. (2004). Control of a two-dimensional movement signal by a noninvasive brain-computer interface in humans. *Proceedings of the national academy of sciences*, 101(51), 17849–17854.

[46] Yao, L., Baker, J. L., Ryou, J.-W., Schiff, N. D., Purpura, K. P., & Shoaran, M. (2020). Mental fatigue prediction from multi-channel ecog signal. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 1259–1263).: IEEE.