COMPARING MEASURES OF BRAIN COMPLEXITY IN THEIR ABILITY TO DISTINGUISH BETWEEN THE RESTING AND PSYCHEDELIC STATE

Candidate number: 229540

Word count: 10, 313

Date: 16/08/2022

Supervisor(s): Dr. Lionel Barnett & Dr. Warrick Roseboom

School of Mathematics and Physical Sciences

University of Sussex

Contents

Abstract

Disorders of decreased consciousness, such as coma and the vegetative state, are characterised by decreased complexity of spontaneous neural signals. On the other end of the spectrum, the psychedelic state shows increased complexity, suggesting this is a state of increased consciousness. Complexity is often computed using the Lempel-Ziv (LZc) compression algorithm, however this measure is flawed. A new measure called CSER has been developed to combat the issues associated with LZc. The current study compares these measures in their ability to measure and distinguish between conscious states (specifically the normal waking and psychedelic states). LZc, CSER and a third discretised entropy measure were calculated from MEG recordings of individuals in a placebo and psychedelic condition. Data was evaluated using exploratory data analysis, significance testing, correlations and machine learning. Both measures found significant differences in complexity between the placebo and psychedelics conditions (W=0.0, p=0.000). Correlations between all three measures were very good (>. 96). The CSER measure produced a better classifier than both LZc and discrete entropy rate; with alterations the classifier produced an accuracy of ~70%. Only the CSER measure showed differences between condition complexity when data was split up by frequency band. Drug type analysis showed that the greatest difference in complexity between conditions are observed in the LSD group. Lastly, sampling variations revealed the prime sampling rate is between 200-300Hz for all measures. CSER most likely produced the best classifier due to its amplification of differences between conditions. Further studies investigating CSER should 1) Assess its correlation with subjective ratings and 2) Evaluate the performance of a classifier fitted with frequency band limited data.

Acknowledgements

I would like to thank a number of people for their support of this dissertation. Firstly, I would like to thank the academics at Imperial College for providing me with the datasets used in the current study. I would also like to thank my supervisors Dr. Lionel Barnett and Dr. Warrick Roseboom whose knowledge of the subject area helped formulate an initial research question, as well as offering suggestions and guidance at all stages. I would also like to that Dr. Lionel Barnett for providing the MATLAB scripts which facilitated the computation of measures in the current paper.

Declarations

This report is submitted as part requirement for the degree of Data Science MSc at the University of Sussex. It is the product of my own labour except where indicated in the text / acknowledgements. The report may be freely copied and distributed provided the source is acknowledged. I hereby give permission for a copy of this report to be loaned out to students in future years.

Signed: Emma Crawford

Introduction

Consciousness:

The empirical study of consciousness' can be traced back over a century to the work of William Wundt (McLeod, S., 2008). Wundt earnt the title 'father of psychology' due to his lab's segregation of psychological study from its philosophical roots. He focused on objectively measuring the components of conscious experiences, such as thoughts and sensations, through a method known as introspection (McLeod, S., 2008). Introspection involves examining and recording one's own internal thoughts and feelings (Flanagan et al, 2015). Although Wundt's work was ground-breaking, introspection was later discredited by the behaviourists, who believed that it was far too subjective. They insisted that only phenomena which can be directly observed and measured are scientifically rigorous, meaning the study of mental processes halted for decades (Flanagan et al, 2015). The cognitive revolution of the 1960's brought the field back to life; psychologists realised that inferences could be made about mental processes based on controlled lab studies (Flanagan et al, 2015). The emergence of brain imaging techniques in the late 1900's was another profound leap for field, allowing psychologists and neuroscientists to directly observe the neural basis of conscious processes (Flanagan et al, 2015) and even correlate these with subjective experiences. This is the main technique used in conscious study today.

A significant amount of research into consciousnesses focuses on investigating 'states of consciousness'. There are four main states in which consciousness is altered: Comatose, the vegetative state, the minimally conscious state and of course sleep (NHS, 2022). Research into these states aims to investigate the characteristic neural patterns and profiles in an attempt to further our understanding of consciousness (Bruno et al. 2011). It has been found

that these decreases in consciousnesses display specific changes in the prevalence of certain frequency bands. When speaking of frequency bands, this is referencing the 5-frequency bands within which spontaneous neural activity occurs in the brain: gamma (32-100Hz), beta (13-32Hz), alpha (8 – 13Hz), theta (4-8Hz), and delta (0.5-4Hz). Higher frequency bands (e.g. gamma or beta) have the shortest wavelengths with more cycles per second, and the lower frequencies (e.g. theta and delta) have longer wavelengths. It has been suggested that high prevalence of longer wavelengths indicates less brain activity and therefore a lower state of consciousness (Purves et al., 2001), and vice versa. This is supported by sleep study findings that light sleep is theta dominant (Purves et al., 2001) and deep sleep delta dominant (Purves et al., 2001), both recognisable by slower wavelengths than the alpha dominant waking brain. The lowest state of consciousness (coma) is also predominately characterised by very low frequency delta waves (<1Hz) (Schiff et al., 2014), whereas the minimally consciousness state, where patients show some outward signs of consciousness, is associated with theta waves (Schiff et al., 2014). In addition to these findings, studies also show that decreased consciousness is associated with decreased complexity of spontaneous neural signals compared to the normal waking state (Casali et al., 2013; Sara & Pistoia et al. 2010). Decreased complexity here means that brain activity is more predictable in reduced states of consciousnesses. As complexity has been found to be a marker of conscious state, its ability to assess and diagnose disorders of consciousness is currently being studied (Górska et al., 2021) and could lead to improved diagnostic capabilities (Coleman et al, 2009), making this an important area of research.

For many years, the waking state has been considered the height of consciousness. However, recent studies have found that spontaneous neural signals display increased complexity after the consumption of psychoactive substances (Schartner et al., 2017;

Timmermann et al., 2019), leading some to conclude that the psychedelic state is one of increased consciousness. Not only does this finding corroborate with anecdotal tales of 'increased awareness' (Gaddum & Vogt, 1956), but it has also been found that the extent to which brain signal complexity increases is correlated with the strength of subjective experience (Schartner et al., 2017). Study of the psychedelic state offers a unique advantage over studies into disorders of consciousnesses in so far as this ability to verify the subjective with the objective. The degree to which complexity increases occur has been found to depend on which drug has been administered. Broadly speaking, there are two classes of psychedelic drug: classic hallucinogens and dissociative drugs. A recent study by Schartner et al (2017) found a drug in the later class (Ketamine) to produce larger differences in complexity compared to two drugs in the former class (LSD and Psilocybin). More research into psychedelics will undoubtedly further our understanding of the relationship between brain signal complexity and conscious level.

Measuring complexity:

The use of signal complexity as a measure of conscious state is closely linked to information theory the idea of the entropic brain (Carhart-Harris, 2014). The entropic brain hypothesis states that the entropy of spontaneous brain activity indicates the 'richness' of a conscious state (Carhart-Harris, 2018). Entropy is a measure which indexes uncertainty about the future behaviour of a dynamical system and the content of future information (Carhart-Harris, 2018). Measures of complexity are used to approximate brain entropy and have been found to do so with high levels of convergence (Amigo et al., 2004; Carhart-Harris, 2018).

Lempel-Ziv complexity:

A popular method for measuring complexity uses the LZc algorithm (Lempel & Ziv, 1976). The algorithm was originally created in 1976 to compress data and therefore reduce the amount of space needed to store it. Although LZc was not created with brain signal analysis in mind, nor was it created with the intent of entropy estimation, it has been hugely successful at doing so (Amigo et al., 2004; Carhart-Harris, 2018). Data passed into the LZc algorithm must first be discretised into two or more values as continuous values cannot be passed in. Data is then divided into substrings by the algorithm based on observed patterns and stored in a dictionary (Figure 1). The number of substrings stored in this dictionary is indicative of the complexity of the data passed into it, with a greater number of dictionary entries meaning that the signal is more complex. A complex signal containing many unique substrings would be very hard to predict the future behaviour of and therefore you can see how this ties in with the definition entropy.

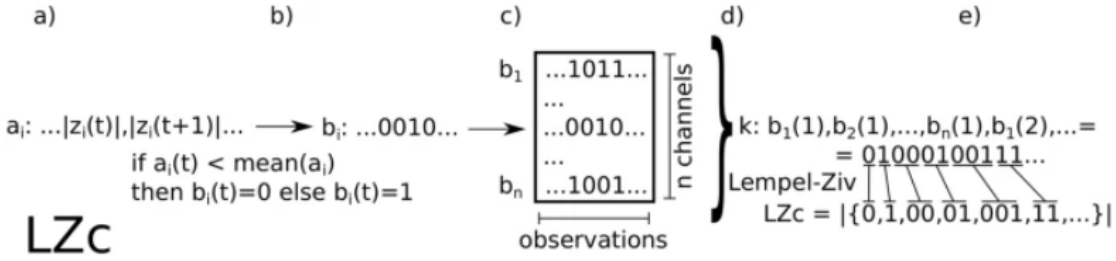


Figure 1. Schematic of the computation of Lempel-Ziv (Schartner et al., 2017)

LZc gained popularity due to its simplicity and fast computation (Kempa et al. 2013) however it is not without its drawbacks, with a recent paper referring to it as a 'flawed tool in neuroscience' (Barnett, L. Private communication, 2022). The main issue being that it requires large windows of data to converge well with entropy rate

and therefore has limited temporal resolution. For studies that wish to investigate very fast acting brain events, such as evoked responses, LZc would be an inappropriate measure of complexity. This means that LZc is limited to the analysis to 'states' only. It would be beneficial to have one complexity measure that is able to cater to all study types as this would allow for more accurate comparisons of brain signal complexity between studies. A second issue with LZc is information loss. Whilst the compression algorithm itself is hailed as 'lossless' (Tank, 2011), the initial discretising step causes large amounts of information to be lost. A measure which can deal with raw values would be better to avoid this.

Complexity via state space models:

A new measure which solves both these issues is complexity via state space models (CSER). Additionally, because the measure is one of predictability, not compressibility like LZc, it resonates more closely with the definition of entropy. As the name suggests, CSER uses state-space modelling (Durbin, 2012) (Figure 2) to estimate entropy rate. Entropy rate can be defined as:

$$H(X) = \lim_{n\to\infty} H = (x_t \mid x_{t-1} \dots \dots . x_2, x_1) \quad \text{Eq. 1}$$

State-space models rely on the premise that some observed data (X) can be modelled as noisy observations of a hidden process (Z). The hidden state Z_t contains all the information needed to predict X_t which is the complete past history of X until time X_{t-1}. Therefore, it follows that state-space entropy rate can be defined as:

$$H(X) = (x_t \mid z_t) \quad \text{Eq. 2}$$

The type of state-space model used in CSER can be represented as a pair of linear equations (Barnett, L. Private communication, 2022)**:**

$$Z_{t+1} = Az_t + K_{\epsilon t}$$

$$x_t = Cz_t + \epsilon_t$$

Eq. 3

Where A is the state transition matrix, C the observation matrix and K the Kalman gain matrix. $\epsilon_t = x_t - \mathbb{E}\{x_t | x_{t-1} \dots \dots . x_2, x_1\}$ and $z_t = \mathbb{E}\{y_t | x_{t-1} \dots \dots . x_2, x_1\}$. As normally distributed noise is assumed in the model, CSER is then calculated with the formula for the entropy of a Gaussian distribution (Cover & Thomas, 2006)**:**

$$h(\mathrm{X}) = \frac{1}{2} \log \det(2\pi e \Sigma) \qquad \textit{Eq. 4}$$

CSER solves the issue faced by LZc of requiring large windows of data to converge with entropy rate because it is able to use past information to calculate instantaneous entropy rates as opposed to requiring very large strings of data. This method also solves the issue of information loss by allowing real values to be passed in.

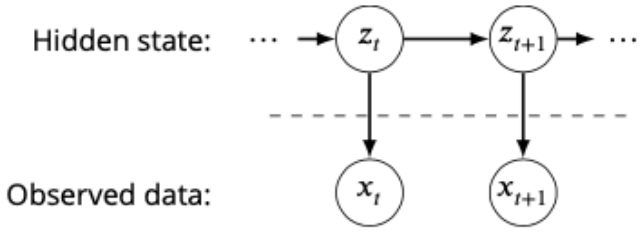


Figure 2. The state-space model (Barnett, L. Private communication, 2022)

Current study:

The current study aims to compare the commonly used LZc measure (Li & Mashour, 2019; Schartner et al. 2017; Scott & Carhart-Harris, 2019) with the new CSER measure in their abilities to distinguish between a placebo and psychedelic state when applied to Magnetoencephalography (MEG) data. The study will also compare these to a discrete entropy rate in order to assess their parity with entropy. For the main analysis, I will firstly use exploratory data analysis methods to ascertain whether measures are capable of detecting differences between Placebo and Drug conditions as well as whether any participants show unusual results. Statistical testing will then be applied to establish whether observed differences are significant or not. Correlational analysis will be run to investigate the level of agreement between the three measures. Lastly, I will train three logistic regression classifiers, one for each of the three measures, in an attempt to predict whether a particular complexity / entropy score belongs to the placebo or psychedelic state. Further analysis will firstly investigate the effect of splitting MEG data into its 5 frequency bands before applying the same analysis as above. The aim of this is to see whether any one of the bands produce greater complexity differences than others**.** I will also briefly investigate differences in complexity between drug types as well as whether varying the sampling frequency when computing measures has any effect on results.

This will further past research in two main ways, firstly through application of the new CSER method and secondly by applying a complexity-based classifier to psychedelic data. Furthermore, sampling frequency alterations were not applied in the similar Schartner et al (2017) paper, so the current study will add to this by implementing such alterations. By applying the new CSER method, the current study aims to prove that this measure is equally as equip as the LZc measure in distinguishing between states of consciousnesses as well as

displaying good agreement with entropy rate. By applying a classifier to the psychedelic data, the current study aims to show that machine learning can successfully be used to distinguish between waking and 'higher states of consciousness' as has already been found for reduced states of consciousness (Wielek et al, 2018).

Hypotheses:

Based on evidence that LZc estimates entropy well (Amigo et al., 2004; Carhart-Harris, 2018) coupled with initial empirical findings using CSER (Barnett, L. Private communication, 2022)**,** I predict that both measures will display an overall significant difference between the drug and placebo conditions, representing increased complexity in the drug condition. I also predict that some channels will display a greater difference between drug and placebo conditions than others and these channels will be in frontal brain regions (Hensler, J., G., 2012). This will be assessed for all three measures using the 10 channels which show the greatest difference between conditions, henceforward referred to as the 'top 10' channels. I also predict that correlations between all three measures will be very high ($r > .90$), as they should be estimating the exact same thing. Due to the shortcomings of LZc (Barnett, L. Private communication, 2022; Zhang et al., 2016) as discussed above, I predict that models trained with LZc data will perform worse on all classification metrics than models trained with CSER data. I predict that when analysing by frequency band, there will be a decrease in complexity between conditions in the alpha band (Schartner et al., 2017) and increases in complexity between conditions in the beta and gamma bands (Purves et al., 2001), which surpass the broadband difference. I also predict, based on findings from Schartner et al (2017), that Ketamine will produce the largest difference between conditions, followed by LSD and Psilocybin will show the smallest differences. Lastly, I predict that sampling frequency variations will have no effect on direction of complexity change,

significance or correlations between measures. However, I do predict an effect on the logistic regression model; when sampling frequency is higher, classification accuracy will be lower (Hounslow et al., 2019) as a result of noise effecting the data.

Data

About the data:

The data used in the current study was provided by Imperial College. There were two sets of data, one containing MEG recordings from 48 participants in a psychedelic drug condition and the other contained MEG recordings form the same 48 participants in a placebo condition, making this study a within-subjects design. This design type was chosen because individuals' resting spontaneous brain activity differs substantially and therefore there is a need to compare the same individual in order to obtain accurate results. The psychedelics administered to participants were Lysergic Acid Diethylamide (LSD), Ketamine (KET) and Psilocybin (PSI); each participant only took one of these. All participants gave informed consent and there were strict exclusion criteria based, see (Muthukumaraswamy et al., 2013; Muthukumaraswamy et al., 2015; Carhart-Harris et al., 2016) for full details on this. All drugs were administered intravenously, with 10-minute MEG recordings following immediately afterwards for KET and PSI and 4 hours later for LSD to capture its peak effect (Stanford Children's Health, 2022). All MEG recordings were high pass filtered at 1Hz, down sampled to 600Hz and divided into 2 second epochs. Epochs containing muscle movement artefacts were removed from the datasets using hanning windowed Fourier transformations to calculate a threshold value for removal. Independent component analysis was also performed to remove any further muscle artefacts, as well as ocular and cardiac

artefacts and this was done to ensure that any body movements would not confound results. The artefact removal process means that the number of epochs varies from participant to participant. Data as pre-processed by Imperial college comes in the form of a three-dimensional array, containing information about channels, observations and epochs. There are 90 source channels in total, all sampled at 600Hz, meaning each 2 second epoch contains 1200 observations.

Computation of measures:

The MATLAB code for calculating my two discrete complexity measures (LZc and discrete entropy) as well for the differential entropy rate (CSER) was downloaded from GitHub (Barnett, 2020; Barnett, 2020). The start-up scripts were adapted to work on my machine, run with the MEG data from Imperial college as input, and the output files saved to my machine. This code output 1 file per participant containing a complexity value per epoch for both drug and placebo conditions. The current study set out to use a sampling rate of 600Hz (as used in Schatner et al., 2017) however, the code for computing differential entropy rate contained a bug when sampling at 600Hz; therefore, I decided to down sample to 300Hz. This is unlikely to have any major effect on analysis as the main difference between 300-600Hz is noise (Barnett, L. Private Communication, 2022b). If anything, a lower sampling rate will improve results, as stated in my hypothesis. Furthermore, because this study is investigating states, not evoked responses, timing is unimportant and therefore a high sampling rate is unnecessary (MNE, 2017).

Data pre-processing:

Complexity / entropy MATLAB output files were read into a python environment (Jyupter notebooks) in which the rest of the analysis took place. A comprehensive Data

Frame was constructed for each of the 48 participants containing all the information that would be required to carry out the analysis. Each dataframe entry included the complexity / entropy value, the condition (Drug vs. Placebo), the type of measure (CSER, LZc or Discrete entropy rate), participant number, drug type (LSD, PSI or KET) and lastly the channel from which that value had been recorded from. Once these were constructed, they were output as CSV files and saved to my machine.

Methods and Results

Preliminary Analysis:

Data was analysed on both an individual level and global level. The reason for analysing individuals separately is because basal brain complexity varies hugely between people (Jaušovec & Jaušovec, 2010; Glahn et al., 2010). Secondly, analysing participants individually allows for the identification of individuals who display interesting or unexpected results, who might go unnoticed otherwise. Data was also analysed with all participant data together (by concatenating the individual data frames), to ensure that patterns observed on the individual level held true at the global level. Bar graphs were plotted to check for direction and magnitude of complexity change between the drug and placebo conditions for the three measures. The majority of participants (44 out of 48) displayed increases in the expected direction, with the drug condition displaying a higher mean complexity compared to the placebo condition for all three measures. Exceptions to this finding were KET participants 6 and 13 as well as PSI participants 12 and 14 who all showed higher complexity during the placebo condition for all three measures (Figure 3). When looking at results using all

participant data, mean complexity is also higher in the drug condition than the placebo condition for all measures (Figure 4).

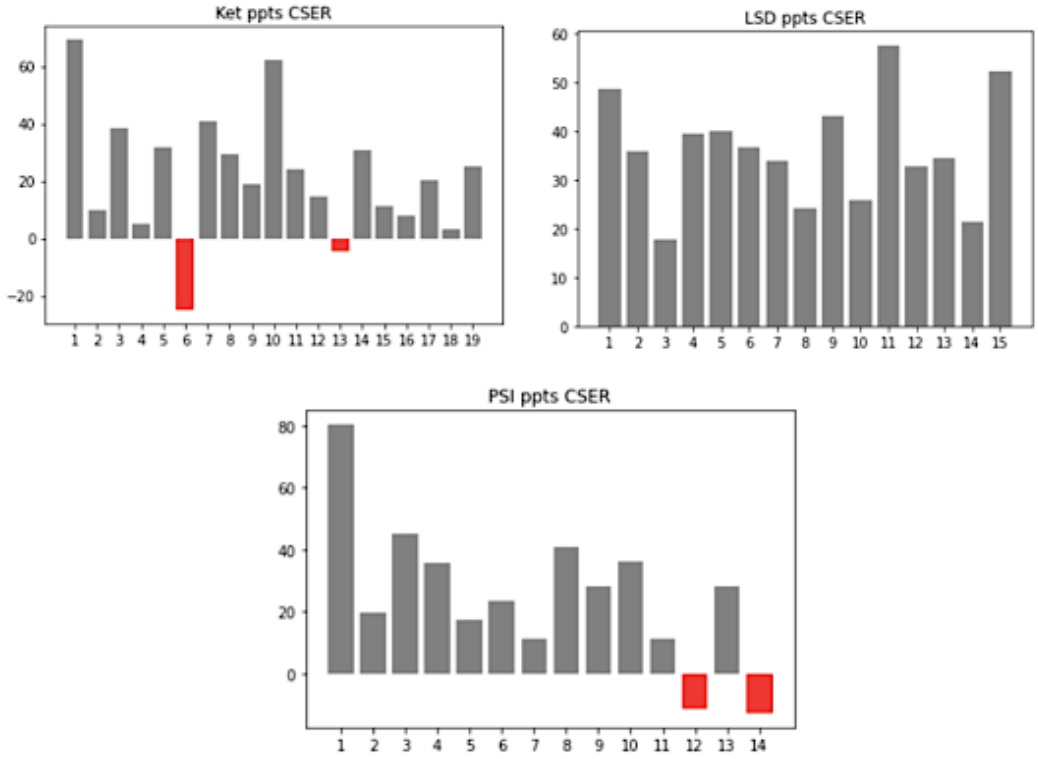


Figure 3. Difference in complexity between placebo and drug conditions for each participant. Grey bars represent complexity being higher in the drug condition and red bars higher in the placebo condition.

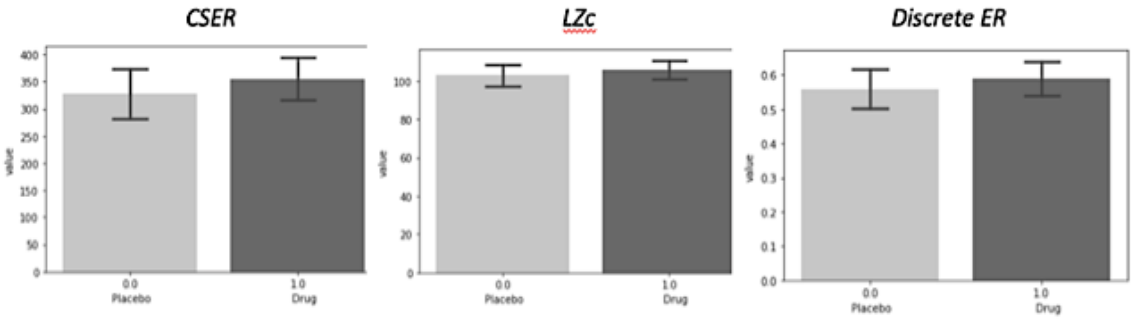


Figure 4. Mean complexity for the drug and placebo conditions for all three measures across all participants

The scales of all three measures differ massively from one another, making direct comparisons between them difficult. Data normalisation was used in an attempt to rescale the data and therefore improve comparability between the measures. However, this seemed to significantly distort my data. Figure 4 clearly shows a far more pronounced difference between conditions for the CSER measure in comparison to LZc, but when data is normalised this difference disappears (Figure 5). For this reason, I decided not to use normalised data in any further analysis. As normalised data could not be used, instead percentage changes were calculated between drug and placebo conditions for all participants and measures. I chose to use percentage change because it is not affected by varying scales of measurement, allowing for fair comparison. Using all participant data, CSER displayed a percentage change in complexity between the conditions of +8.4%, Discrete entropy rate +5.5% and LZc the lowest change at +2.9%. This shows that CSER is amplifying differences between conditions and that LZc may be underestimating the difference between conditions, when comparing the to the discrete entropy rate. This may have an impact on the result of significance testing.

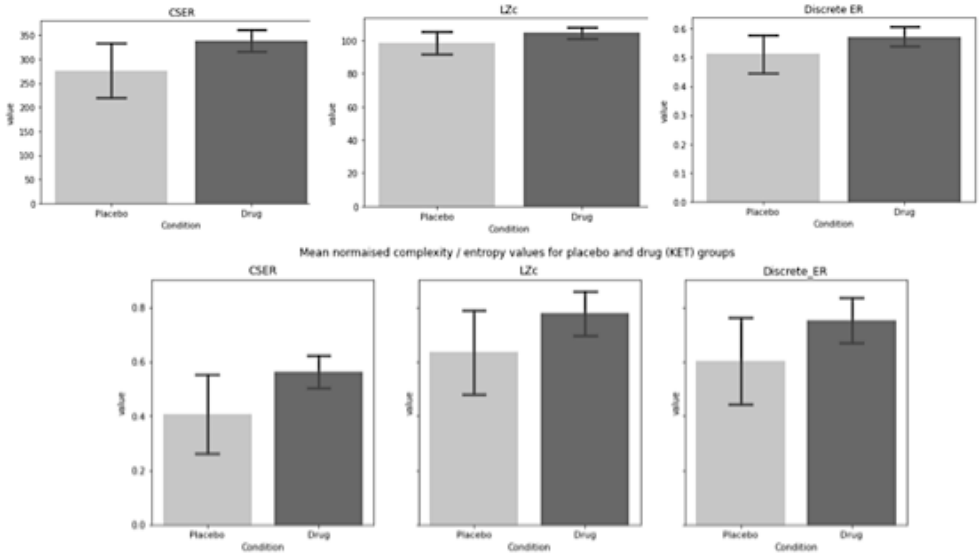


Figure 5. Distortion of data with normalisation, Bar graphs on the top row show differenced plotted for each measure on their own scale. Bar graphs on the bottom show normalised plot. Example taken from KET ppt 10.

Before analysis of the 'top 10' channels could be done, value differences for all 90 channels first needed to be calculated. For each channel, mean complexity / entropy was computed for both drug and placebo conditions. Following this, I computed the value difference between these conditions by subtracting the mean placebo value from the mean drug value. This process was repeated from all three measures. Table 1 displays the 'top 10' channels for each measure (and the lobe of the brain they are located in); value difference was used to ascertain this. Out of the 14 channels that appear in Table 1 just 1 is located within the frontal lobe, 2 within the parietal lobe, 4 in the temporal lobe and 7 lie within the occipital lobe. With only 14 channels appearing in total, this is a sign of good agreement between measures.

Top 10 channels

CSER	Lempel-Ziv	Discrete Entropy Rate
Caneus_R (Occipital)	Caneus_R (Occipital)	Calcarine_L (Occipital)
Calcarine_L (Occipital)	Calcarine_L (Occipital)	Lingual_L (Occipital)
Lingual_L (Occipital)	Lingual_L (Occipital)	Caneus_R (Occipital)
Occipital_Sup_R (Occipital)	Temporal_Mid_L (Temporal)	Temporal_Mid_L (Temporal)
Precuneus_R (Parietal)	Precuneus_R (Parietal)	Temporal_Sup_R (Temporal)
Occipital_Sup_L (Occipital)	Temporal_Sup_R (Temporal)	Precuneus_R (Parietal)
Cuneus_L (Occipital)	Cuneus_L (Occipital)	Heschl_R (Temporal)
Clacarine_R (Occipital)	Clacarine_R (Occipital)	Occipital_Sup_R (Occipital)
Temporal_Mid_L (Temporal)	Occipital_Sup_R (Occipital)	Frontal_Inf_Oper_L (Frontal)
Postcentral_R (Parietal)	Heschl_R (Temporal)	Temporal_Sup_L (Temporal)

Table 1. Top 10 channels identified using value difference scores for CSER, LZc and Discrete entropy rate. Channels are presented form greatest difference (top) to smallest (bottom).

Significance testing:

The Wilcoxon Signed rank test:

The Wilcoxon signed rank test is a non-parametric test for significance. The null hypothesis (that there is no difference between condition means) will be rejected only if the W statistic is less than or equal to some critical value (LaMorte, 2017). This critical value will be very high for the current study because there are a very large number of observations in the current dataset. The W statistic is calculated using the following formula:

$$W = \sum_{i=1}^{N_r} [sgn(x_{2,i} - x_{1,i}) \cdot R_i] \qquad \textit{Eq. 5}$$

Equation 5 calculates the difference between pairs of scores, in our case, between drug (x_2) and placebo (x_1) scores for each channel (N = 90). The capital R in Equation 5 stands for rank; difference values are ordered from lowest to highest (with the sign ignored) and then given a rank from 1 to N, depending on where they fall (LaMorte, 2017). The sign of the differences is then applied to the rank score. Signed ranks are summed to give us our W statistic. If this statistic falls below some critical value, we can reject the null hypothesis and conclude that there is a real difference between conditions. The critical value differs for every experiment as it depends on sample size and alpha level (LaMorte, 2017). The reason a Wilcoxon test was selected for this study is because the study follows a repeated measures design (participants took part in both drug and placebo conditions). Additionally, histograms show violation of the assumption of normality in the data (Appendix 1), meaning a non-parametric test is more powerful in this scenario.

Wilcoxon tests were performed on an individual and whole group level using data from all three measures; these were implemented using SciPy's Wilcoxon signed-rank test (Virtanen et al, 2020). For the most part, all three measures showed a significant difference at $\alpha = .05$, whereby mean complexity / entropy was higher in the drug than the placebo condition. Exceptions to this were KET participants 4 and 18 who showed a non-significant difference on all three measures. Additionally, KET participants 6 and 13 and PSI participants 12 and 14 all showed a significant difference in the unexpected direction, with complexity in the placebo condition significantly greater than in the drug condition. When all participant data was used together, the result is a significant difference in the hypothesised direction for all three measures ($W=0.0$, $p = 0.000$).

Correlations:

Correlation is a measure which reveals the extent to which two variables are linearly related, in the case of this paper, it will be used as a measure of agreement between the three complexity / entropy measures. Seaborn heatmaps were created to visually display correlations between the three measures. The value differences per channel (as calculated in the preliminary analysis section) obtained using each of the three measures were the values passed into heatmaps. On the individual participant level, all but 3 participants showed very high correlations between measures, these were PSI participants 5 and 12 and LSD participant 6. When using all participant data together, correlations between all measures were very high (> .96) (Figure 5). LZc and discrete entropy rate show a perfect positive correlation with one another ($r = 1$), LZc and CSER have the next highest correlation coefficient ($r = .97$) and lastly discrete entropy rate and CSER show the worst, but still a very good correlation coefficient of $r = .96$.

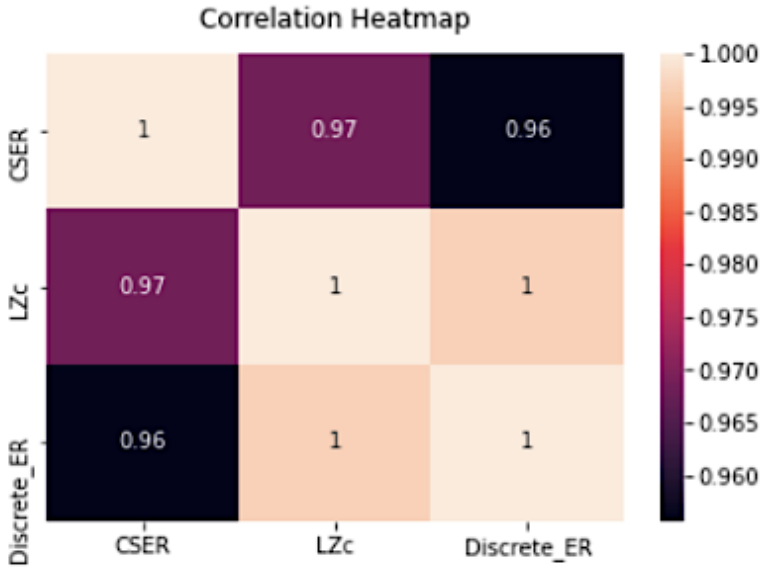


Figure 5. Correlations between the three measures of complexity / entropy over all 90 source channels

Logistic regression classifier:

How it works:

Logistic regression models learn a mapping from inputs to classes, using conditional probability. These probabilities are obtained using the logistic sigmoid function:

$$\sigma(x) = \frac{exp^x}{1 + exp^x} \qquad Eq.6$$

This function takes the linear transformation term ($W^T X + b$) and transforms it to the [0,1] probability space:

$$\sigma(x) = \sigma(W^T X + w_0) \qquad Eq.7$$

The aim is to learn which function $f(x)$ maps our data X to the [0,1] output space best. This is achieved by the logistic regression algorithm learning which weights and bias terms create the best fitting function. Optimal weights are those which provide the maximum likelihood of getting our correct training labels (e.g. Drug or Placebo):

$$\arg\max_{w} \sum_{n=1}^{k} \log(\sigma\,(y_n W^T X_n)) \qquad \textit{Eq.8}$$

In addition to this, our optimal weights will also be those which minimise log loss:

$$\arg\min_{w} \sum_{n=1}^{k} -\log(\sigma\,(y_n W^T X_n)) \qquad \textit{Eq.9}$$

Depending on the exact algorithm used to perform logistic regression, log loss is minimised either with gradient or co-ordinate decent. In this paper, I will be using the liblinear solver meaning it will be minimised with co-ordinate decent, this iteratively updates just one weight at a time until there are no more updates to be made (aka. Convergence) (Johnson & Guestrin, 2017).

I have chosen a supervised learning approach over an unsupervised learning model (such as K-means clustering) due to the fact that boxplots show significant overlapping between drug and placebo data (Appendix 2), making it very difficult for the algorithm to split the data points into two groups accurately. I have chosen to use a logistic regression classifier over a more complex model (such as a neural network) because on very large datasets, such as the one in the current paper, a simpler model will facilitate much faster computation. Additionally, simpler models are less likely to overfit (Bouso, 2015). Logistic regression models are also common practice for binary classification tasks (Singhal, 2020) such as the one I am faced with (Drug Vs. a

Placebo). Having said this, if the logistic regression model performs very poorly, a neural network may also be tested.

Metrics:

The current study will use 5 metrics to evaluate model performance: Accuracy, Precision, Recall, F1 score, and Area Under the Receiver Operating Curve (AUC-ROC). Accuracy, also referred to as 'hit rate' is simply a measure of the number of correct classifications divided by the total number of observations (Bagheri, 2019):

$$Accuracy = \frac{TP+TN}{Total\ number\ of\ observations} \qquad Eq.\ 10$$

TP here stands for True positive and TN true negative. This metric is often misleading (Medium, 2019), hence the need for additional metrics. Precision is a measure of how many of the observations which were predicted as being in the positive class are truly positive (Bagheri, 2019):

$$Precision = \frac{TP}{TP+FP} \qquad Eq.\ 11$$

FP here stands for false positive. Recall, sometimes referred to as 'sensitivity' or 'true positive rate' is another commonly used measure to assess binary decision classifiers. Recall is a measure of how many of the observations from the positive class actually get predicted as being positive (Bagheri, 2019):

$$Recall = \frac{TP}{TP+FN} \qquad Eq.12$$

FN here stands for false negative. Whilst both informative, recall will be more influential than precision when assessing models because in the context of the current study it bares more importance. This is because, if the model were to go on to be used to assess disorders on consciousness' for example, wrongly thinking somebody is less conscious than they truly are could lead to the decision to end a life wrongly. However, if you were to incorrectly predict somebody as being in a higher state of consciousness than they truly are, whilst this might lead to poor precision, the real-life consequences are not so bad. A metric used to find the mean of both precision and recall is the F1 score (Korstanje, 2021):

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall} \qquad Eq.13$$

Lastly, because I know that data in the current study is fairly balanced between drug and placebo labels and because the decision is a binary one, I can also use AUC-ROC metric. An ROC curve plots true positives against false positives (Narkhede, 2018). The area under this curve provides a score which reflects the model's ability to distinguish between classes (Narkhede, 2018) and is calculated using integration (Aim, 2020):

$$AUC - ROC = \int_0^1 TPR(FPR)\ dFPR$$

$$= \int_0^1 TPR(FPR^{-1}(x))\ dx \qquad Eq.14$$

TPR (True positive rate) here refers to the metric ‘recall’ and FPR (False positive rate) is a measure of how many positive predictions are truly negative (Bagheri, 2019):

$$FPR = \frac{FP}{TN+FP} \qquad Eq.15$$

Model training and evaluation:

Before classifiers could be run a few pre-processing steps were required. Firstly, the condition labels of ‘drug’ and ‘placebo’ were converted to the numeric values of ‘1’ and ‘0’ retrospectively, in order to be accepted by the classifier. Following this, data was randomly split into train and test sets of size 70% and 30% retrospectively. Lastly, because data is not normally distributed, normality was asserted using a log transformation of X values. Models contained only one input feature, which was the complexity / entropy value, with which to predict drug or placebo labels from. During model training, 5-fold cross validation was used (Pedregosa et al., 2011) to reduce the chances of the model overfitting to the training data, allowing for better generalisation at test (Pramoditha, 2021). 5-fold cross validation works by splitting the training data into a further 5 random sections. Then 5 separate models are fitted, with each model using K-1 folds for training, and the additional fold for validation. The fold used for validation differs for each of the 5 models, thereby so does the combination of training folds. This variation of folds used to train the model is key to how overfitting is avoided. The cross validated accuracy score is an average of accuracy scores over all 5 models. Figure 6 shows a schematic of how 5-fold cross validation works.

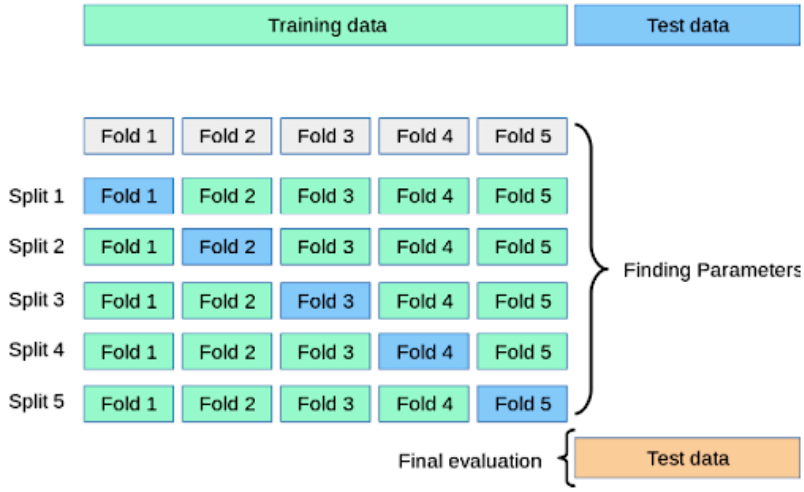


Figure 6. 5-fold cross validation (Pedregosa et al., 2011)

Cross validated logistic regression models (Pedregosa et al., 2011) were trained on the combined data of all 48 participants. 3 models were trained in total, using values from each of the three complexity / entropy measures. When these models were evaluated on test data (Figure 6), CSER outperformed the other two models on 4 out of the 5 metrics and as a result produced around 8 - 10 thousand less misclassifications than them (Appendix 4). Although recall is .001 lower in the CSER model than the LZc model, this is negligible. Metrics barely differed between train (Appendix 3) and test (Figure 6), meaning there are no signs of overfitting. As the CSER model performed better than LZc and Discrete ER models, the rest of this section will focus on improving CSER based models only. Both Random Forest and Neural Network classifiers were fitted the data because 1) They have shown success in a similar study (Wielek et al., 2018) and 2) They may display improved performance. However, these more complex models showed negligible differences and took much longer to run, therefore the rest of this section will focus solely on logistic regression models.

CSER

Metric	Value
Accuracy	.631
Precision	.612
Recall	.667
F1 Score	.631
AUC - ROC	.632

Lempel-Ziv

Metric	Value
Accuracy	.613
Precision	.592
Recall	.668
F1 Score	.612
AUC - ROC	.614

Discrete ER

Metric	Value
Accuracy	.617
Precision	.599
Recall	.651
F1 Score	.616
AUC - ROC	.617

Figure 6. Test metrics for CSER, LZc and Discrete entropy rate models (using all 48 participants)

Removal of outliers:

Although the CSER model was able to predict drug and placebo labels substantially above chance level, performance is inadequate for usage as a diagnostic tool in a clinical setting, therefore I sought to make improvements. A common technique for improving model performance is outlier removal (Ray, 2015). Preliminary analysis identified four participants who displayed higher brain complexity values in the placebo condition rather than the drug condition. Data points from these participants is likely having a negative impact on model performance, so they were removed from the dataset. Data was again split into train (70%) and test (30%) and X values log transformed. As before, a cross validated logistic regression was trained and evaluated on test data (metrics in Table 2). As you can see, the model now performs significantly better than before, with all metrics jumping up 1 – 1.8%. In terms of misclassifications this means the model is making ~24,000 less

misclassifications than before, this was estimated with the help of confusion matrices (Appendix 4; Appendix 5).

Metric	Value
Accuracy	.647
Precision	.630
Recall	.677
F1 Score	.647
AUC - ROC	.648

Table 2. Test metrics for CSER classifier after outlier removal

Reduction of Channels:

When conducting preliminary analysis, 10 channels were identified which exhibited the greatest difference between drug and placebo conditions. It has already been established that CSER produces a better classifier because it inflates differences between drug and placebo conditions. Following on from this logic, designing a classifier which only uses channels which inflate differences between drug and placebo conditions, should also improve model performance. Another logistic regression classifier was trained using the top 10 channels as identified by the CSER measure (Table 1). Test metrics from this model can be seen in Table 3. Comparing these this to Table 2, increases of between 2.7 – 3.1% are observed for the five metrics.

Metric	Value
Accuracy	.678
Precision	.657
Recall	.704
F1 Score	.677
AUC-ROC	.678

Table 3. Test metrics when logistic regression model is trained only with the 'top 10' CSER channels.

Metric	Value
Accuracy	.695
Precision	.679
Recall	.720
F1 Score	.695
AUC-ROC	.695

Table 4. Test metrics when logistic regression model is trained only with the 'top 3' CSER channels

Following the success of this, another classifier was fit, using just the top 3 channels identified using the CSER measure (Table 1). Table 1 shows that the top 3 channels are identical for all three measures, meaning they must sit quite reliably above the other channels. Test metrics for this classifier are displayed in Table 4. This show a further increase in all 5 metrics by 1.6 – 2.2%. Accuracy, F1 Score and AUC-ROC have all reached a respectable 70% and Recall 72%. For a visual representation

of how this model performs in comparison to a random and perfect classifier as well as in comparison to the original.

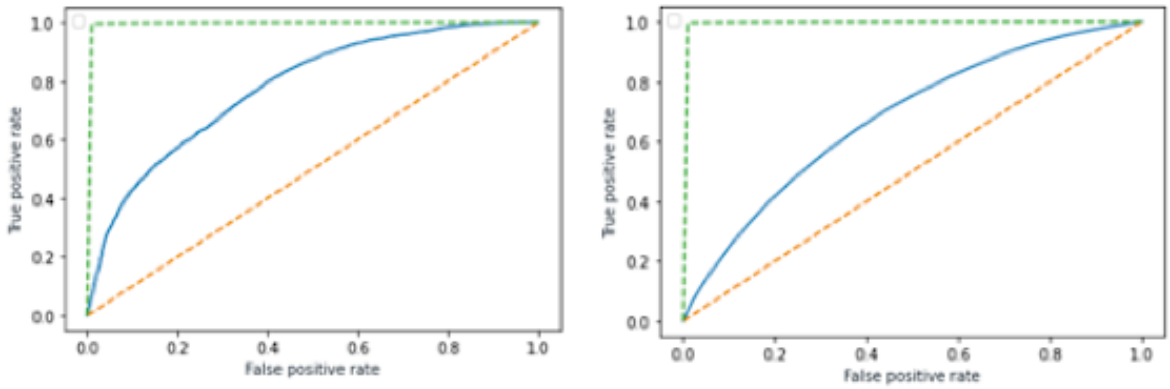


Figure 7. Left ROC curve shows CSER classifier when only the top 3 channels were used. Right ROC curve shows original unaltered model. Blue lines represent the model, orange dotted lines a random classifier and green dotted lines a perfect classifier.

Lone participant model performance:

Whilst classification has been significantly improved, metrics are still questionable for clinical usage. Preliminary analysis in the current study as well as prior research (Jaušovec & Jaušovec, 2010; Glahn et al., 2010) indicates that basal brain complexity differs massively from person to person, meaning that one individuals drug complexity value could be equal to another's complexity placebo value, regardless of whether there was an increase between conditions or not. This is undoubtedly negatively affecting the classifiers ability to label values as 'drug' or 'placebo' when all participant data is used together. Therefore, a logistic regression model was fit using data from just one individual (KET participant 1) to see how performance improves. Test metrics and the ROC curve for this model can be seen in Figure 8. You can see the model performs much better, however, using data from just one participant, has serious implications for model applicability; this will be examined in the discussion section.

Metric	Value
Accuracy	.819
Precision	.803
Recall	.828
F1 Score	.819
AUC - ROC	.819

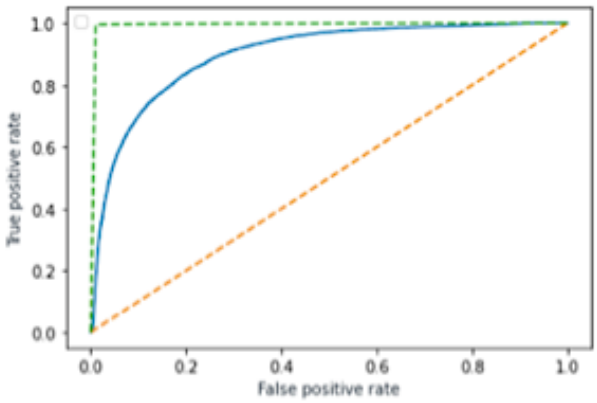


Figure 8. Test metrics and ROC curve for logistic regression model fitted on data from KET participant 1 only.

Further Analysis

Frequency band analysis:

Unfortunately, code to band-limit the new CSER method was still in development until very recently meaning only rudimentary comments will be made regarding frequency band differences based on the CSER measure. For all three measures, MATLAB scripts were run five more times, once for each of the five frequency bands: delta, theta, alpha, beta and gamma. These frequency bands are selected for in the scripts using bandpass filtering. Scripts were, as before, run with a sampling rate of 300Hz. LZc and Discrete entropy rate output was loaded into python and dataframe made for each participant with 5x complexity values per epoch, one for each band. Bar plots show very little discrepancy between frequency bands when it comes to complexity differences between drug and placebo conditions (Appendix 6). In support of this, correlational analysis revealed perfect agreement (r = 1) between the five frequency bands (Appendix 7), suggesting that complexity behaves the same

across all bands in our dataset when using LZc and Discrete entropy rate as measures. For this reason, it would not be worth performing single band classifiers (e.g. an alpha or theta only classifier) as it would not improve classification beyond what has already been achieved. Significant tests revealed that differences in complexity between drug and placebo conditions were significant for all five frequency bands (W=0.0, p = 0.000), when using both LZc and Discrete entropy rate. Lastly, agreement between frequency bands regarding which channels increase in complexity the most with psychedelic intake was very good; for the most part the top 10 channels were identical.

Although bands show no distinction in complexity difference across the two conditions, Figure 9 shows that absolute complexity values do differ between frequency bands, with some starting higher (beta) and some lower (gamma) than others. Therefore, a logistic regression classifier was fit containing 5 separate features for each epoch; these were the complexity values obtained from each of the five frequency bands. In this case, increasing the feature space should improve classifier performance as bar graphs indicate that these bands can be considered as somewhat distinct features from one another. For the classifier fir using LZc data improvements of 1.9 – 2.7% were seen across all five metrics when compared to the identical single feature LZc model (Table 5). For the mode fit using Discrete entropy data, slightly larger gains of 2.4 – 3.3% were observed across metrics when compared to its single feature counterpart (Table 6).

LZc single feature model		LZc 5 feature model	
Accuracy	.628	Accuracy	.655
Precision	.608	Precision	.642
Recall	.676	Recall	.695
F1 Score	.628	F1 Score	.655
ROC-AUC	.629	ROC-AUC	.655

Table 5. Comparison of metrics between the single feature and the 5-feature model using LZc data

Discrete ER single feature model		Discrete 5 feature model	
Accuracy	.631	Accuracy	.656
Precision	.613	Precision	.642
Recall	.664	Recall	.697
F1 Score	.631	F1 Score	.655
ROC-AUC	.632	ROC-AUC	.656

Table 6. Comparison of metrics between the single feature and the 5-feature mode using Discrete entropy rate data

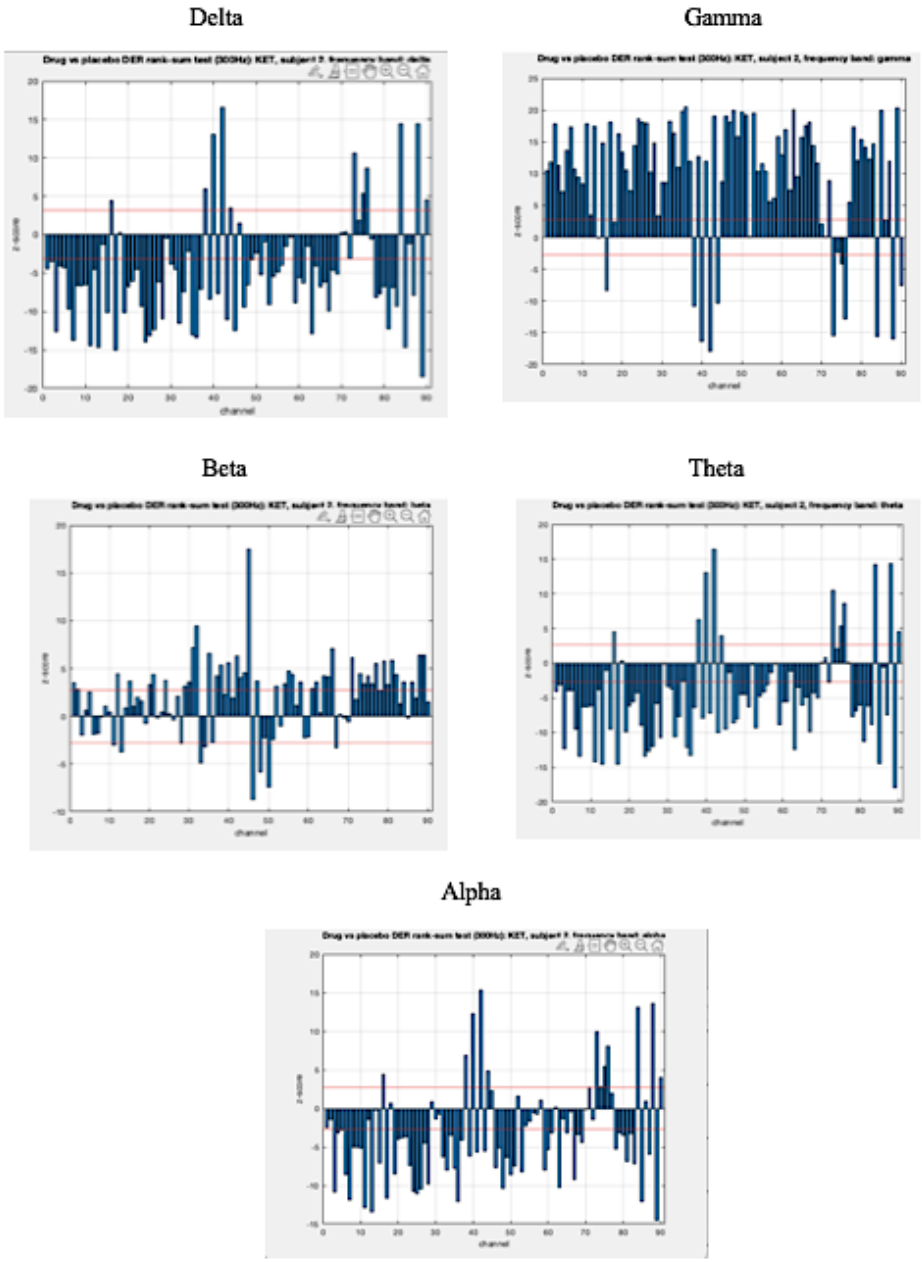


Figure 9. Plots showing complexity difference between drug and placebo conditions for each of the 5 frequency bands when using the CSER measure. Positive values represent an increase in complexity in the drug condition and vice versa. Example taken from KET participant 2.

As mentioned, frequency band scripts were run for the CSER measure, but time constraints did not allow for further analysis in python. The MATLAB code used to run CSER measure computation also comes with a statistics routine, which outputs a figure displaying Z-Score differences between conditions for the 90 channels. This routine was run for a handful of participants in an attempt to identify a pattern. Results showed that delta, theta and alpha band complexity all exhibited an overall decrease in complexity between drug and placebo conditions and that beta and gamma, an overall increase. See Figure 9 for an example from KET participant 2.

Drug type analysis:

When making the original participant data frames, information on drug type was included, meaning that analysis on drug differences simply required sub sectioning these existing data frames before running analysis. All three measures were in agreement that LSD produced the greatest complexity difference between conditions (Table 7). LZc and Discrete entropy measures found KET to show greater complexity differences between conditions than PSI (Table 7). CSER on the other hand found PSI to show greater complexity differences than KET (Table 7). The difference between mean complexity in the drug and placebo conditions reached significance for all three drug types across all three measures ($W=0.0$, $p= 0.000$). Mean complexity, as before, referring to the mean complexity value across epochs for each of the 90 channels.

Channels showing the greatest complexity differences between conditions varied substantially between the three drug types (Appendix 8) but was fairly consistent across measures. This finding is supported by the fact that correlations showed very poor agreement between drug types for all three measures (Appendix 9). Correlations were weakly negative

between LSD and KET (LZc = -.14, CSER= -.09, Discrete ER = -.13) and also between PSI and LSD (LZc = -.19, CSER= -.11, Discrete ER = -.22). Correlations were weakly to moderately positive between KET and PSI (LZc = .29, CSER= .46, Discrete ER = .24). When looking at the 10 channels with the greatest complexity differences (top 10 channels), for LSD these lay primarily in the frontal and temporal lobes. For KET, primarily in the occipital and parietal lobes and PSI the temporal and occipital lobes. This again supports the findings above, that these three drugs are exerting their influence on different channels to one another.

Measure	Sampling rate	% change between conditions
CSER	KET	6.8%
CSER	LSD	10.9%
CSER	PSI	8.9%
LZc	KET	2.3%
LZc	LSD	4.1%
LZc	PSI	2.1%
Discrete ER	KET	4.3%
Discrete ER	LSD	7.7%
Discrete ER	PSI	4.1%

Table 7. Percentage change between placebo and drug conditions for each drug type and each measure.

Logistic regression classifiers were fit with data form each of the three drugs. Models trained with PSI data performed worst for all three measures, with KET models performing very slightly and LSD data producing significantly improved models when compared to all

drug models (Figure 10). CSER LSD models actually outperformed all models thus far in the current paper (expect for the lone participant model).

LZC

All data model		PSI model	KET model	LSD model
Accuracy	.628	.592	.606	.689
Precision	.608	.566	.593	.679
Recall	.676	.525	.607	.736
F1 Score	.628	.591	.606	.688
ROC-AUC	.629	.588	.606	.688

Discrete ER

All data model		PSI model	KET model	LSD model
Accuracy	.631	.593	.609	.693
Precision	.613	.571	.592	.683
Recall	.664	.502	.632	.738
F1 Score	.631	.590	.609	.692
ROC-AUC	.632	.587	.610	.692

CSER

All data model		PSI model	KET model	LSD model
Accuracy	.647	.601	.619	.728
Precision	.630	.578	.601	.719
Recall	.677	.529	.641	.761
F1 Score	.647	.600	.619	.727
ROC-AUC	.648	.597	.619	.727

Figure 10. Metric for model produced using each of the three drug types, for all three measures

Varying Sampling frequencies:

Thus far, all measures have been computed using a sampling frequency of 300Hz (300 observations per second of data). All data was both up sampled to 600Hz and down sampled to 200Hz and 150Hz. As with frequency band analysis, this involved re running MATLAB scripts with the new sampling frequencies. Sampling at 600Hz had an unusual effect on the CSER measure, whereby values became negative for both drug and placebo

conditions (Figure 11). This is an effect that high levels of noise can have on the CSER measure. This meant that when interpreting results, a less negative value in the drug group compared to placebo was considered as an increase in complexity.

I found that as sampling rate increases, so does the difference between conditions (Table 8). CSER produces a particularly large increase when the sampling rate is 600Hz compared with the other two measures. All sampling rates showed significant differences between groups (W=0.0, p=0.000). When looking at correlations between measures, they were very high (r > .95) for all four sampling frequencies. When fitting models with the CSER measure, I found that data sampled at both 200 and 300Hz performed in equal standing to one another and performed best out of all sampling frequencies. 150Hz models performed between 1.3 – 2.9% worse across metrics than the 200 and 300Hz measures. Although the 600Hz sampling rate far outperformed other sampling rates in percentage change between conditions (Table 8), when it came to classification, the 600Hz model performed significantly worse, between 4.2 – 4.8%. Using LZc, resulted in again the best performance being jointly shared by 200 and 300Hz models and 600Hz models only performed slightly worse. The 150Hz model performed worst of all four models by 1.8 – 2.7% across metrics, except recall which was actually improved. For models fit with discrete entropy rate, the 300Hz model performed best, 200Hz slightly worse, 600Hz slightly worse than that and 150Hz the worst of all four. Full metrics can be found in Appendix 10.

Measure	Sampling rate	% change between conditions
CSER	150Hz	5.3%
CSER	200Hz	6.4%
CSER	300Hz	8.4%
CSER	600Hz	30.2%
LZc	150Hz	1.5%
LZc	200Hz	1.9%
LZc	300Hz	2.9%
LZc	600Hz	3.8%
Discrete ER	150Hz	3.5%
Discrete ER	200Hz	4.1%
Discrete ER	300Hz	5.5%
Discrete ER	600Hz	6.4%

Table 8. Percentage changes between drug and placebo conditions by measure type and sampling rate

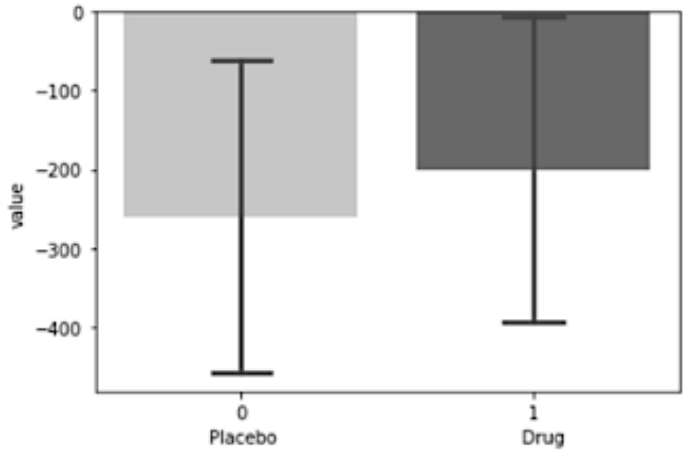


Figure 11. CSER results when sampled at 600Hz

Discussion

The main aim of the current study was to compare the new CSER measure to the commonly used LZc measure (Li & Mashour, 2019; Schartner et al. 2017; Scott & Carhart-Harris, 2019) in their ability to distinguish between a psychedelic (KET, LSD and Psilocybin) or placebo brain state as well as explore if there were any conditions under which this ability was affected. The two measures were additionally compared with a third Discrete entropy measure throughout, to test their entropy estimation accuracy. Another major aim of the current study was to apply machine learning methods to complexity measures using psychedelic data for the first time.

Explanation of main findings:

As hypothesised, both measures detected significant differences between drug and placebo conditions at $\alpha = .05$, representing higher brain complexity in the drug than the placebo condition. Although CSER and LZc both detected significant differences, bar graphs (Figure 4) and percentage change calculations indicate that the CSER measure causes differences between conditions to be more pronounced than does the LZc measure. As the discrete entropy measure fell roughly between the two in terms of discrimination between conditions (Figure 4), it suggests LZc may underestimates entropy, and CSER overestimate it. Therefore, LZc could be suited to more stringent studies as differences are less likely to be found, and CSER to less stringent studies. I am unsure why this finding occurred, but interesting nonetheless.

When analysing participant data individually, results that were both non-significant and significant in the incorrect direction were found. There could be a number of reasons

why participants show non-significant results, one being drug tolerance. Drug tolerance is a reduced reaction to a drug following its repeated use, leading to larger and larger quantities being required to produce the same effects (Lynch, 2022). There is little information provided by Imperial college on the history of participants psychedelic use, apart from that participants had all tried them at least once before (Schartner et al., 2017). Studies show that LSD (Buchborn et al., 2016), PSI (Passie et al., 2002) and KET (White & Ryan, 1996) are all prone to tolerance issues and they are all popular recreational drugs in today's society (Orhurhu et al., 2019; Halberstadt et al., 2019; Fricke, 2019). Making drug tolerance a highly plausible explanation for non-significant results. Again, there could be a number of reasons why participants displayed the opposite relationship to the hypothesis, whereby complexity was higher in the placebo condition. One example is the placebo effect; the placebo effect is a beneficial effect produced by a drug which has no real therapeutic value. Therefore, the effect can be attributed a person's belief in the drug (Oxford Languages, 2022). Placebo's have been found to produce objective changes to physiological functioning sometimes even 'exceeding those attributed to potent pharmacologic action' (Harrington, 1999). If a participant really believed they were taking part in the psychedelic condition when in fact they were in the placebo condition, this mere belief could be enough to flip results.

As hypothesised, some channels displayed greater differences between drug and placebo conditions than others. Contrary to my hypotheses, the top 10 channels identified were primarily in the occipital lobe. I predicted that frontal lobe channels would show the greatest difference as this is where 5-HT_{2A} receptors are most abundant; these receptors are thought to be the mechanism through which psychedelics exert their effects (Madsen et al., 2019). One reason for these unexpected findings could simply be that activity does not equal complexity. Just because activity increases most in frontal regions (as a result of 5-HT_{2A}

activation), it does not automatically mean this activity will be the most complex / unpredictable. Secondly, my hypothesis failed to consider that KET actually acts via NMDA recptors which are distributed equally throughout the entire brain (Mion & Villevieille, 2013; Dzamba et al., 2013).

As hypothesised, correlations between all three measures were very high (r = .96 – 1.00). CSER obtained the lowest correlations with other measures (LZc: .97; Discrete ER: .96), whilst LZc and Discrete entropy had a perfect correlation (1.00), suggesting that LZc might estimate entropy better than CSER. Having said this, the similarities between LZc and Discrete entropy could be attributed to them both going through the same discretisation procedure before computation, which CSER did not. CSER may in fact be truer to the entropy of the raw MEG values. Further research should compare CSER to a non-discretised entropy measure, such as permutation entropy (Matilla-García & Marín, 2008). Three participants showed far worse correlations between measures than the others (r = .72 - .84). 2/3 of these participants present no other abnormal results, making it unclear what has caused this.

Lastly, as hypothesised, it was found that models trained with CSER data outperformed LZc trained models and since formulating this hypothesis, the reason behind why has become clearer. Because CSER amplifies differences between conditions, it is easier for the decision boundary to separate data points belonging to the two classes as they are physically further apart from one another. Models were tweaked to improve performance in three ways; the first being outlier removal. Data points from participants who displayed higher complexity in the placebo condition were removed (Figure 1). The idea behind this was that removing data points exhibiting the reversed relationship between condition and

complexity value to the rest of the dataset should improve performance, which it did. The second tweak was reduction of channels used in the classifier. When only the top 3 CSER channels were used, accuracy and F1 scores reached .695. This is because these channels amplify differences between conditions again meaning it is easier for the decision boundary to separate data points belonging to the two classes. Lastly, A model was built using data from only one participant. This was done because individuals resting brain complexity can vary substantially (Shumbaywonda et al., 2018; Hornero et al., 2010, Shumbaywonda et al., 2019) As would be expected, results were massively improved. This is because when all participant data is used together, one person's placebo complexity value may be the same as another's drug complexity value, irrespective of a complexity increase / decrease between conditions. This therefore makes single subject classification a much simpler task.

Explanation of further findings:

In addition to the main analysis, further analysis was performed to test whether there were any conditions under which the above findings would be different. For LZc and Discrete entropy, full analysis was re-run with complexity values split up into by frequency band. Due to time constraints, only a brief analysis was run for the 5 frequency bands with CSER. Contrary to my hypothesis, LZc and Discrete entropy found no difference between frequency bands in terms of complexity increases from placebo to drug condition. Interestingly, when using the CSER measure, results appeared as expected. Different band pass filtering methods are adopted in the discrete measures and CSER MATLAB scripts. The choice of band pass filtering method could be the reason LZc and Discrete entropy rate are displaying conflicting result to previous studies (Schartner et al., 2017) and the CSER measure. Logistic regression models were fit using the 5 frequency bands to make a 5 feature in classifier, which improved classification metrics. Improvements were observed because

absolute values varied between bands, making them distinct features from one another, and therefore useful additions to the classifier.

Drug analysis yielded unexpected results. Contrary to predictions, LSD displayed the greatest positive difference between drug and placebo conditions and as a result produced the best classifier of all three drugs. Results from Schartner et al (2017), which uses the same data as the current study, found that KET produced the most pronounced difference between conditions. One reason for this discrepancy could be the way in which LZc has been calculated in the two studies. Many variations of the LZc algorithm exist (MIT, 2022) all calculating it in slightly different ways. Schartner et al (2017) use phase randomised renormalized LZc calculations in many of their diagrams, which they themselves found to have an imperfect correlation with the standard LZc measure. Further investigation will need to take place to corroborate or falsify differences observed in the current paper. I also found that correlations between drug types were very poor, suggesting that these drugs produce their effects via different channels in the brain (supported by the figure in Appendix 8). This was supported by findings that the top 10 channels from each of the three drugs varied substantially, eluding to differing pharmacological mechanisms of action between the drugs. Correlational analysis also revealed that KET and PSI show much better correlations with one another when the CSER measure is used. Suggesting that data discretisation might be losing some important complexity information for either KET or PSI data.

In the final section of my further analysis, I varied the sampling frequencies at which measures were calculated. I noticed that with decreased sampling frequency, came decreased percentage changes between conditions. Lower sampling frequencies lead to lower absolute complexity values, however, percentage changes between conditions should remain

unaffected by this as both conditions are sampled at the same rate. Further investigation is required into why this occurs. The CSER measure was significantly affected by a sampling rate of 600Hz. Negative values and very large standard deviations (Figure 11) suggest that CSER displays a high level of noise when sampled at 600Hz. This is supported by the finding that even though CSER showed massive differences between conditions (30.2%), the classifier performed the worst of any in this entire paper. This is because high noise models perform poorly on tasks of classification (Gupta & Gupta, 2019). Although LZc and Discrete entropy were less obviously affected by noise at 600HZ, classifiers using 600Hz data still performed worse than those trained with 200 – 300Hz data. 150Hz classifiers performed the worst of all four for LZc and Discrete entropy, suggesting that a sampling rate of 200 – 300Hz is the sweet spot between too much noise, and not enough data to accurately estimate complexity / entropy.

Real world applications:

One application of the current study is to the diagnosis of conscious states, in particular disorders of consciousness'. A previous study has managed to apply LZc to the discrimination of unresponsive wakefulness syndrome and the minimally conscious state with relative success (Górska et al., 2021). The current study shows that the CSER classifier is able to discriminate between conscious states better than the LZc classifier, meaning that this new measure could be a really useful tool in the diagnosis of disorders of consciousness above and beyond models that are already being investigated.

A second application of the new CSER method is in academic research into the complexity of brain states. The current study has confirmed that the measure correlates very highly with entropy (r = .96), which is the aim of a good complexity measure. The fact that

CSER can be used across a wider range of study types than LZc (as referenced in the introduction) makes it an attractive option. Having such a diverse complexity measure would allow for better comparability of results across complexity studies.

One last application of the current study is in the treatment of disorders of consciousness. This study confirms the findings of others (Schartner et al., 2017; Timmermann et al., 2019) that brain complexity increases under the influence of psychedelic drugs, using three different measures. Therefore, there is now enough evidence to suggest that psychedelics could be used to raise complexity in those with DOC's (which are characterised by decreased complexity) with the hope of reinstating normal consciousness. A recent study (Schmitz et el., 2022) suggests that the feasibility of such a treatment may depend on the individual patient, this is because individuals react very differently to psychedelics as a result of genetic differences. This finding is supported by my investigations at the individual participant level, where I discovered noticeable differences between individuals.

Study problems and limitations:

Following on from the idea of using the CSER measure to classify and diagnose disorders of consciousness', whilst performance of the top 3 channels CSER model was good (classification accuracy of ~70%), there is no guarantee that the same method would work for disorders of consciousness. The primary issue here being that, the channels most effected by disorders of consciousness are highly likely to be different than those most effected by psychedelic drugs. Research into which channels these are would need to be done before implementing any classifier. The lone participant classifier actually performed even better (classification accuracy of ~80%) with no need for channel reduction. Whilst this

performance is impressive, using data from just one participant is problematic for applicability. If the model were to be used to diagnosis a DOC, a lack of access to data from the patients' normal waking brain would be a major issue.

A second issue, which applies to the field of conscious study as a whole, is the idea of 'conscious level'. The current study, and many before it, rely on the premise that consciousness' falls into discrete categories (e.g. the vegetative or minimally conscious state). However, in reality, healthcare professionals often struggle to diagnose between such disorders (Fischer & Truog, 2017) due to unclear distinctions. One recent book suggests that the cause of such misdiagnosis' stems from the fact that consciousness actually falls on a continuum (Fischer & Truog, 2017). In the current study, the extent to which complexity increased in the psychedelic state varied massively between individuals, with some showing huge differences and some only very small. This evidence supports ideas that consciousness falls on a continuum because individuals within the same conscious state varied tremendously. Forcing my data into discrete categories could be why classification algorithms struggled to surpass 70% on whole participant data.

Bigger picture questions:

One question this paper set out to address is whether or not LZc is a useful tool in its current usage (studies of conscious states). LZc has shown consistent results across a plethora of studies including those investigating sleep (Pregowska et al.; Li & Wang, 2010), disorders of consciousnesses' (Górska et al. 2021) and the psychedelic state (Schartner et al., 2017; Farnes et al., 2020) as well as behaving as expected in the current study. LZc can also be used with a variety of brain imaging techniques such as EEG (Farnes et al., 2020), fMRI (Hudetz et al., 2016) and MEG (Schartner et al., 2017). Whilst CSER is a promising alternative to

LZc for low noise data, at high levels of noise the measure was profoundly affected, whereas LZc was not. The next step in investigating whether CSER could be a more useful measure than LZc would be to compare the both to a non-discretised entropy rate. If LZc shows correlations with this equal to CSER, it would show that information loss does not affect LZc's ability to accurately estimate entropy and indicates that LZc is a useful and sufficient measure for studies of conscious state.

The second question this paper wanted to address was whether the word 'complexity' is a suitable term for LZc and CSER. This will be done by comparing complexity in the current study to complex systems. A complex system is often characterised by what is known as 'spontaneous order' or 'self-organisation'. This means that although compents of the system may not appear all that ordered, the overall result of interactions between them is robust order (UoB, 2022). Several examples of this can be observed in nature; the building of termite mounds, the shoaling of fish and flocking of birds are all good ones as they display this highly ordered global behaviour (UoB, 2022). Complexity in the sense of the current study is an estimation of entropy, with higher entropy / complexity actually meaning that the brain is functioning in a more disordered and unpredictable fashion. This does not resonate with the notion of complex systems that has just been described above. Therefore, perhaps the term complexity is not the best way to describe what LZc and CSER are really measuring.

The final question this paper will address is whether LZc and CSER estimate entropy well. In short, complexity and entropy are not the same thing but, complexity does estimate entropy very well to a certain extent. As illustrated in Figure 12, entropy and complexity follow the same trajectory up until complexity reaches a plateau and begins decreasing. Therefore, when entropy gets larger, and the system is very disordered, the two measures can

actually become highly divergent from one another. The complexity measures used in the current study however were both found to estimate entropy rate well. LZc and CSER both showed very strong correlations with entropy rate (> .96) and for the most part, results obtained using the two measures were very similar to those obtained from the discrete entropy measure.

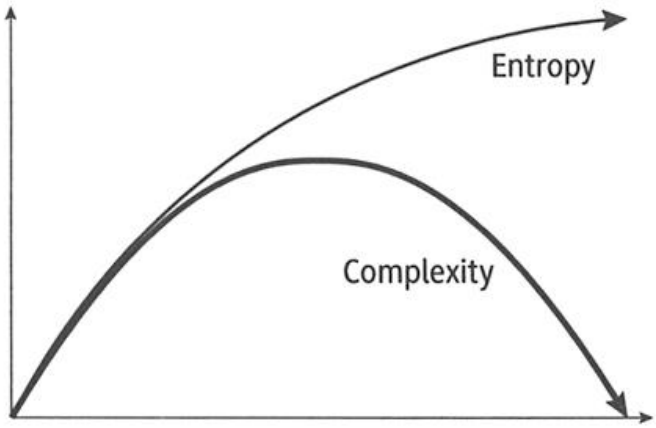


Figure 12. Trajectory of complexity and entropy (Jordão et al., 2019)

Future directions:

In the original study, Imperial college also collected subjective ratings from their participants. Unfortunley, investigating the relationship between my three measures and the subjective ratings was not possible within the scope of this study. Schartner et al (2017) found evidence of correlations between complexity and subjective ratings using LZc and various other complexity / entropy measures. It would be great to add to this finding by assessing how well the new CSER measure correlates with subjective ratings, ultimately to evaluate its ability to capture real phenomena. Subjective measures could also be used to test my placebo effect hypothesis as mentioned earlier in the discussion. If participants who showed increased brain complexity in the placebo over the drug condition also display

increased subjective ratings, it would be indicative of the placebo effect. If not, then some other variable must be at play here.

As it stands, much of the literature into conscious states focuses on reduced consciousness, but studies into elevated states could be just as useful in promoting our understanding of the topic. Examples of elevated states are mania and psychosis, which have both displayed increased brain complexity (Bahrami et al., 2005; Sokunbi et al., 2014). Whilst mania and psychosis are typically associated with mental health issues, they could also be considered 'disorders of consciousness', but in an elevated sense. A more solid understanding of where these states fall in relation to the normal waking and psychedelic state would massively improve our understanding of not only consciousnesses but of these mental health conditions themselves. CSER could be a particularly useful tool for investigation into these states given its superior discriminatory power over other methods (as found in the current study).

There are a couple of avenues the current study would have explored if the time frame had allowed and which are definlety worth exploring next. The first would have been to complete a more thorough analysis and modelling of the CSER data when split up by frequency band. Initial results indicated increased complexity in the gamma and beta bands, but decreased complexity in the alpha, theta and delta bands (Figure 9). These were differences that were not detected by LZc or Discrete entropy band limited data. Based on these findings, I would predict that CSER data split up by frequency band will produce a much-improved model compared to that of LZc and Discrete entropy rate, due to the highly distinct nature of each of the CSER frequency bands. Secondly, I would have increased the alphabet size parameter (which was 2 in the current study) when calculating my discretised

measures in MATLAB to see how this affects results. I predict that an increased alphabet size will lead to less data loss; this is because the more categories the raw data can be separated into, the more accurately it will be reflected. This could lead to increased correlations and reduced differences between LZc / Discrete entropy and the lossless CSER measure, hence improving confidence in the new measure.

Conclusion:

To summarise, the current study found that LZc and CSER are both able to detect significant differences between placebo and psychedelic conditions with very good correlations observed between the two measures as well as with a third discretised entropy measure. The new CSER method was able to classify complexity values from placebo and psychedelic groups with greater success than LZc, when logistic regression models were fitted. Models were improved by means of outlier removal and channel reduction, leading to a final classification accuracy of ~70%. Only the CSER measure was able to detect complexity differences between the five frequency bands within the brain; gamma and beta frequencies showed increased complexity between conditions whilst alpha, theta and delta showed decreases across the majority of channels. Drug type analysis revealed that LSD produced the greatest differences between conditions for all three study measures, with an LSD only classifier reaching ~72% accuracy. Sampling frequency variations revealed the best results were produced when sampling between 200-300Hz for all three measures. The current study displays the potential of the new CSER method for measuring the complexity of brain data and distinguishing between states of consciousness. Further studies investigating CSER should 1) Assess its correlation with subjective ratings and 2) Evaluate the performance of a classifier fitted with frequency band limited data.

Words = 10, 313

References

Aim, N., 2020, Mathematics behind ROC-AUC interpretation, https://medium.com/p/e4e6f202a015, Accessed on: 25/07/2022

Amigó, J.M., Szczepański, J., Wajnryb, E. and Sanchez-Vives, M.V., 2004. Estimating the entropy rate of spike trains via Lempel-Ziv complexity. *Neural Computation*, *16*(4), pp.717-736.

Bagheri, R., 2019, ROC Curve, a Complete Introduction, https://towardsdatascience.com/roc-curve-a-complete-introduction-2f2da2e0434c, Accessed on: 25/07/2022

Bahrami, B., Seyedsadjadi, R., Babadi, B. and Noroozian, M., 2005. Brain complexity increases in mania. *Neuroreport*, *16*(2), pp.187-191.

Barnett, L, 2020, fLZc (Version #1) [C]. https://github.com/lcbarnett/fLZc

Barnett, L, 2020, MVGC2 (Version #1) [MATLAB]. https://github.com/lcbarnett/MVGC2

Barnett, L., 2022, Private communication, Spectrally and temporally resolved estimation of neural signal diversity.

Bouso, J.C., Palhano-Fontes, F., Rodríguez-Fornells, A., Ribeiro, S., Sanches, R., Crippa, J.A.S., Hallak, J.E., de Araujo, D.B. and Riba, J., 2015. Long-term use of psychedelic drugs is associated with differences in brain structure and personality in humans. *European Neuropsychopharmacology*, *25*(4), pp.483-492.

Bruno, M.A., Vanhaudenhuyse, A., Thibaut, A., Moonen, G. and Laureys, S., 2011. From unresponsive wakefulness to minimally conscious PLUS and functional locked-in syndromes: recent advances in our understanding of disorders of consciousness. *Journal of neurology*, *258*(7), pp.1373-1384.

Buchborn, T., Grecksch, G., Dieterich, D.C. and Höllt, V., 2016. Tolerance to lysergic acid diethylamide: overview, correlates, and clinical implications. *Neuropathology of Drug Addictions and Substance Misuse*, pp.846-858.

Carhart-Harris, R.L., Leech, R., Hellyer, P.J., Shanahan, M., Feilding, A., Tagliazucchi, E., Chialvo, D.R. and Nutt, D., 2014. The entropic brain: a theory of conscious states informed by neuroimaging research with psychedelic drugs. *Frontiers in human neuroscience*, p.20.

Carhart-Harris, R. L. et al, 2016, Neural correlates of the lsd experience revealed by multimodal neuroimaging. *Proceedings of the National Academy of Sciences* **113**, 4853–4858

Carhart-Harris, R.L., 2018. The entropic brain-revisited. *Neuropharmacology*, *142*, pp.167-178.

Casali, A.G., Gosseries, O., Rosanova, M., Boly, M., Sarasso, S., Casali, K.R., Casarotto, S., Bruno, M.A., Laureys, S., Tononi, G. and Massimini, M., 2013. A theoretically based index of consciousness independent of sensory processing and behavior. *Science translational medicine*, *5*(198), pp.198ra105-198ra105.

Coleman, M.R., Davis, M.H., Rodd, J.M., Robson, T., Ali, A., Owen, A.M. and Pickard, J.D., 2009. Towards the routine use of brain imaging to aid the clinical diagnosis of disorders of consciousness. *Brain*, *132*(9), pp.2541-2552.

Durbin J, Koopman SJ., 2012 Time Series Analysis by State Space Methods. *Oxford University Press.*

Dzamba, D., Honsa, P. and Anderova, M., 2013. NMDA receptors in glial cells: pending questions. *Current neuropharmacology*, *11*(3), pp.250-262.

Farnes, N., Juel, B.E., Nilsen, A.S., Romundstad, L.G. and Storm, J.F., 2020. Increased signal diversity/complexity of spontaneous EEG, but not evoked EEG responses, in ketamine-induced psychedelic state in humans. *Plos one*, *15*(11), p.e0242056.

Fischer, D.B. and Truog, R.D., 2017. *When bright lines blur: deconstructing destinctions between disorders of consciousness*. New York, NY: Oxford University Press.

Flanagan, C., Berry, D., Jarvis, M., Liddle, R. (2015), *AQA Psychology for A-Level*, Cheltenham, Illuminate Publishing.

Fricke, J., Lenz, C., Wick, J., Blei, F. and Hoffmeister, D., 2019. Production options for psilocybin: making of the magic. *Chemistry–A European Journal*, *25*(4), pp.897-903.

Gaddum,. J.H., Vogt, M. (1956), Some central actions of 5-Hydroxytryptamine and various antagonists. Pharmacol, 11, pp.175

Glahn, D.C., Winkler, A.M., Kochunov, P., Almasy, L., Duggirala, R., Carless, M.A., Curran, J.C., Olvera, R.L., Laird, A.R., Smith, S.M. and Beckmann, C.F., 2010. Genetic control over the resting brain. *Proceedings of the National Academy of Sciences*, *107*(3), pp.1223-1228.

Górska, U., Rupp, A., Celikel, T. and Englitz, B., 2021. Assessing the state of consciousness for individual patients using complex, statistical stimuli. *NeuroImage: Clinical*, *29*, p.102471.

Gupta, S. and Gupta, A., 2019. Dealing with noise problem in machine learning data-sets: A systematic review. *Procedia Computer Science*, *161*, pp.466-474.

Halberstadt, A.L., Klein, L.M., Chatha, M., Valenzuela, L.B., Stratford, A., Wallach, J., Nichols, D.E. and Brandt, S.D., 2019. Pharmacological characterization of the LSD analog N-ethyl-N-cyclopropyl lysergamide (ECPLA). *Psychopharmacology*, *236*(2), pp.799-808.

Harrington, A. ed., 1999. *The placebo effect: An interdisciplinary exploration* (Vol. 8). Harvard University Press.

Hornero, R., Ortiz, T. and Lopez-Ibor, J., 2010. Analysis of brain complexity and mental disorders. *Actas Esp Psiquiatr*, *38*(4), pp.229-238.

Hounslow, J.L., Brewster, L.R., Lear, K.O., Guttridge, T.L., Daly, R., Whitney, N.M. and Gleiss, A.C., 2019. Assessing the effects of sampling frequency on behavioural classification of accelerometer data. *Journal of experimental marine biology and ecology*, *512*, pp.22-30.

Hudetz, A.G., Liu, X., Pillay, S., Boly, M. and Tononi, G., 2016. Propofol anesthesia reduces Lempel-Ziv complexity of spontaneous brain activity in rats. *Neuroscience letters*, *628*, pp.132-135.

Jaušovec, N. and Jaušovec, K., 2010. Resting brain activity: differences between genders. *Neuropsychologia*, *48*(13), pp.3918-3925.

Johnson, T.B. and Guestrin, C., 2017, July. StingyCD: Safely avoiding wasteful updates in coordinate descent. In *International Conference on Machine Learning* (pp. 1752-1760). PMLR.

Júnior, Jordão & Santos, Talysson & Rodrigues Mendes Ribeiro, Rafael & Avila, Ivonete & Maciel, Carlos. (2019). Entropy: from thermodynamics to signal processing. 10.17648/sbai-2019-111127.

Kempa, D. and Puglisi, S.J., 2013, January. Lempel-Ziv factorization: Simple, fast, practical. In *2013 Proceedings of the Fifteenth Workshop on Algorithm Engineering and Experiments (ALENEX)* (pp. 103-112). Society for Industrial and Applied Mathematics.

Korstanje, J., 2021, The F1 score, https://towardsdatascience.com/the-f1-score-bec2bbc38aa6, Accessed on: 25/07/2022

LaMorte, W., 2017, *Wilcoxon Signed Rank Test, https://sphweb.bumc.bu.edu/otlt/mph-modules/bs/bs704_nonparametric/BS704_Nonparametric6.html* , Accessed on: 20/07/2022.

Lempel, A. and Ziv, J., 1976. On the complexity of finite sequences. *IEEE Transactions on information theory*, *22*(1), pp.75-81.

Li, D. and Mashour, G.A., 2019. Cortical dynamics during psychedelic and anesthetized states induced by ketamine. *Neuroimage*, *196*, pp.32-40.

Li, L. and Wang, R., 2010, June. Complexity analysis of sleep EEG signal. In *2010 4th International Conference on Bioinformatics and Biomedical Engineering* (pp. 1-3). IEEE.

Lynch, S. S., 2022, *Tolerance and Resistance to Drugs, https://www.msdmanuals.com/en-gb/home/drugs/factors-affecting-response-to-drugs/tolerance-and-resistance-to-drugs* , Accessed on: 25/07/2022

Madsen, M.K., Fisher, P.M., Burmester, D., Dyssegaard, A., Stenbæk, D.S., Kristiansen, S., Johansen, S.S., Lehel, S., Linnet, K., Svarer, C. and Erritzoe, D., 2019. Psychedelic effects of

psilocybin correlate with serotonin 2A receptor occupancy and plasma psilocin levels. *Neuropsychopharmacology*, *44*(7), pp.1328-1334.

Massachusetts Institute of Technology, *Lempel-Ziv.* https://math.mit.edu/~shor/PAM/lempel_ziv_notes.pdf , Accessed on: 15/08/2022

Matilla-García, M. and Marín, M.R., 2008. A non-parametric independence test using permutation entropy. *Journal of Econometrics*, *144*(1), pp.139-155.

McLeod, S. (2008), *Wilhelm Wundt*, https://www.simplypsychology.org/wundt.html, Accessed on: 18/07/2022

Medium.com, 2019, Why "accuracy" is a misleading metric in machine learning, https://medium.com/analytics-vidhya/why-accuracy-is-a-misleading-metric-in-machine-learning-3939fd37b942

Mion, G. and Villevieille, T., 2013. Ketamine pharmacology: an update (pharmacodynamics and molecular aspects, recent findings). *CNS neuroscience & therapeutics*, *19*(6), pp.370-380.

MNE, 2017. *Filtering and resampling data.* *https://mne.tools/0.15/auto_tutorials/plot_artifacts_correction_filtering.html* Accessed on: 19/07/2022

Muthukumaraswamy, S. D. et al. Broadband cortical desynchronization underlies the human psychedelic state. *The Journal of Neuroscience***33**, 15171–15183 (2013).

Muthukumaraswamy, S. D. et al. Evidence that subanesthetic doses of ketamine cause sustained disruptions of nmda and ampa-mediated frontoparietal connectivity in humans. *The Journal of Neuroscience* **35**, 11694–11706 (2015).

Narkhede, S., 2018, Understanding AUC - ROC Curve, https://towardsdatascience.com/understanding-auc-roc-curve-68b2303cc9c5

NHS, 2022, *Disorders of consciousness*, https://www.nhs.uk/conditions/disorders-of-consciousness/, Accessed on: 15/08/2022

Orhurhu, V.J., Vashisht, R., Claus, L.E. and Cohen, S.P., 2019. Ketamine toxicity.

Oxford Languages, 2022, *placebo effect,* *https://www.google.com/search?client=safari&rls=en&q=definition+of+the+placebo+effect&ie=UTF-8&oe=UTF-8* , Accessed on: 26/07/2022

Passie, T., Seifert, J., Schneider, U. and Emrich, H.M., 2002. The pharmacology of psilocybin. *Addiction biology*, *7*(4), pp.357-364.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W.

Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E.A. Quintero, Charles R Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. (2020) SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17(3), 261-272.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. the Journal of machine Learning research, 12, 2825-2830

Pregowska, A., Proniewska, K., van Dam, P. and Szczepanski, J., 2019. Using Lempel-Ziv complexity as effective classification tool of the sleep-related breathing disorders. *Computer Methods and Programs in Biomedicine*, *182*, p.105052.

Purves D, Augustine GJ, Fitzpatrick D, et al., 2001, Neuroscience. 2nd edition, Sinauer Associates, https://www.ncbi.nlm.nih.gov/books/NBK10996/

Ray, S. 2015, 8 Proven Ways for improving the "Accuracy" of a Machine Learning Model, https://www.analyticsvidhya.com/blog/2015/12/improve-machine-learning-results/ , Accessed on: 25/07/2022

Sarà, M. and Pistoia, F., 2010. Complexity loss in physiological time series of patients in a vegetative state. *Nonlinear dynamics, psychology, and life sciences*, *14*(1), p.1.

Schartner, M.M., Carhart-Harris, R.L., Barrett, A.B., Seth, A.K. and Muthukumaraswamy, S.D., 2017. Increased spontaneous MEG signal diversity for psychoactive doses of ketamine, LSD and psilocybin. *Scientific reports*, *7*(1), pp.1-12.

Schiff, N.D., Nauvel, T. and Victor, J.D., 2014. Large-scale brain dynamics in disorders of consciousness. *Current opinion in neurobiology*, *25*, pp.7-14.

Schmitz, G.P., Jain, M.K., Slocum, S.T. and Roth, B.L., 2022. 5-HT2A SNPs Alter the Pharmacological Signaling of Potentially Therapeutic Psychedelics. *ACS Chemical Neuroscience*.

Scott, G. and Carhart-Harris, R.L., 2019. Psychedelics as a treatment for disorders of consciousness. *Neuroscience of consciousness*, *2019*(1), p.niz003.

Singhal, M., 2020, *Logistic Regression*, https://medium.com/@memegha24k/logistic-regression-e9078930ae3a, Accessed on: 15/08/2022

Sokunbi, M.O., Gradin, V.B., Waiter, G.D., Cameron, G.G., Ahearn, T.S., Murray, A.D., Steele, D.J. and Staff, R.T., 2014. Nonlinear complexity analysis of brain FMRI signals in schizophrenia. *Plos one*, *9*(5), p.e95146.

Stanford Children's health, 2022. *All About LSD.* *https://www.stanfordchildrens.org/en/topic/default?id=all-about-lsd-1-2604* Accessed on: 19/07/2022

Tank, M.K., 2011. Implementation of Lempel-ZIV algorithm for lossless compression using VHDL. In Thinkquest~ 2010 (pp. 275-278). Springer, New Delhi.

Timmermann, C., Roseman, L., Schartner, M., Milliere, R., Williams, L.T., Erritzoe, D., Muthukumaraswamy, S., Ashton, M., Bendrioua, A., Kaur, O. and Turton, S., 2019. Neural correlates of the DMT experience assessed with multivariate EEG. *Scientific reports*, *9*(1), pp.1-13.

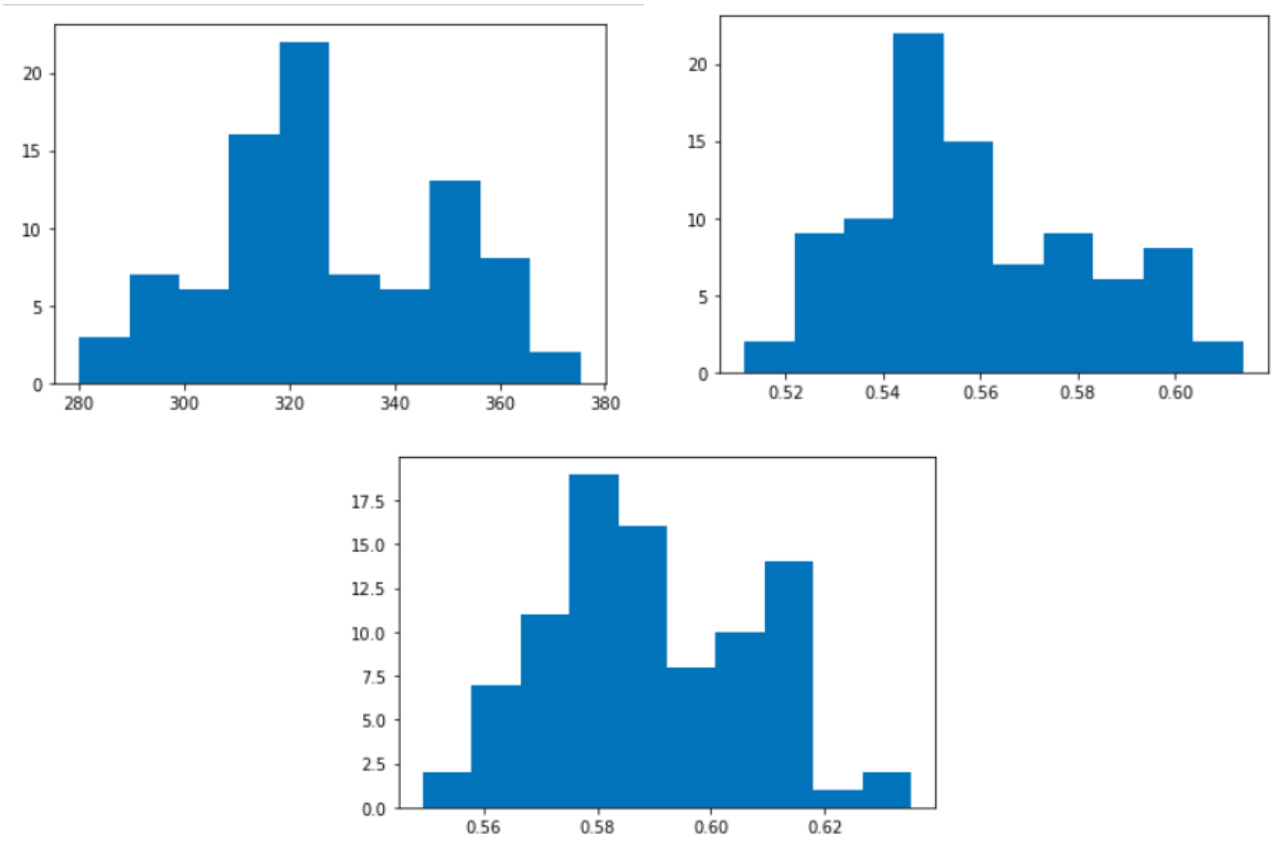
University of Bristol, 2022, *Using reason to define a complex system*, *https://www.bristol.ac.uk/research/impact/defining-complex-system/* , Accessed on 27/07/2022

Wielek, T., Lechinger, J., Wislowska, M., Blume, C., Ott, P., Wegenkittl, S., Del Giudice, R., Heib, D.P., Mayer, H.A., Laureys, S. and Pichler, G., 2018. Sleep in patients with disorders of consciousness characterized by means of machine learning. *PloS one*, *13*(1), p.e0190458.

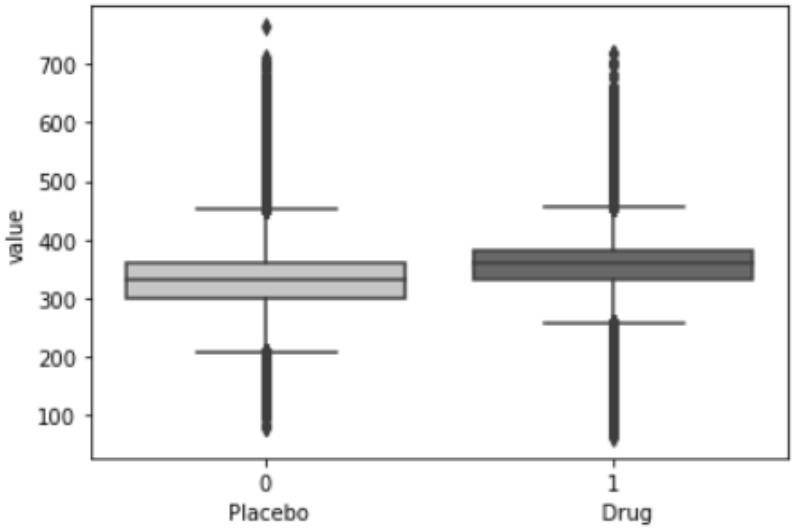
White, J.M. and Ryan, C.F., 1996. Pharmacological properties of ketamine. *Drug and alcohol review*, *15*(2), pp.145-155.

Zhang, Y., Wei, S., Di Maria, C. and Liu, C., 2016. Using Lempel–Ziv complexity to assess ECG signal quality. *Journal of medical and biological engineering*, *36*(5), pp.625-634.

Appendix



Appendix 1. Histograms showing non-normality of data.



Appendix 2. Boxplots showing high levels of overlapping data in drug and placebo conditions

CSER

Metric	Value
Accuracy	.631
Precision	.613
Recall	.667
F1	.631
AUC – ROC	.632

Lempel-Ziv

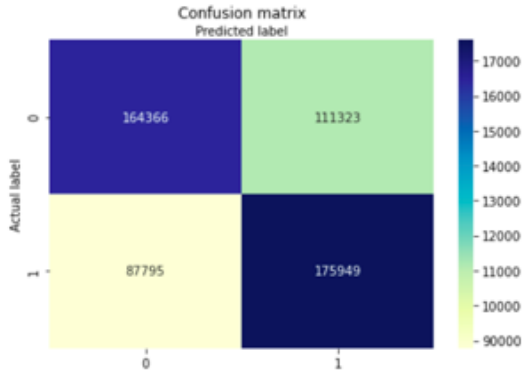
Metric	Value
Accuracy	.613
Precision	.592
Recall	.668
F1	.612
AUC – ROC	.614

Discrete ER

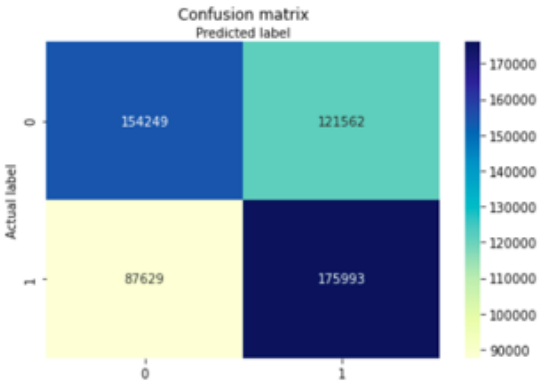
Metric	Value
Accuracy	.616
Precision	.599
Recall	.649
F1	.615
AUC – ROC	.616

Appendix 3. Train metrics for CSER, LZc and Discrete ER models (using all 48 participants)

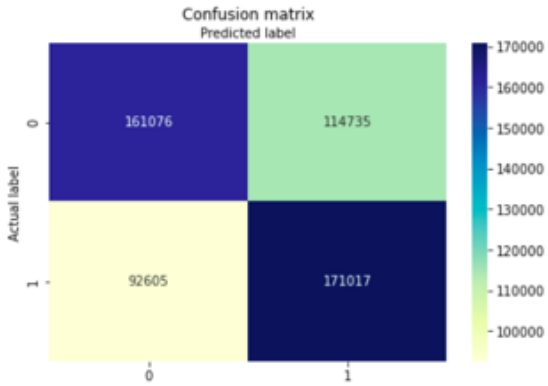
CSER



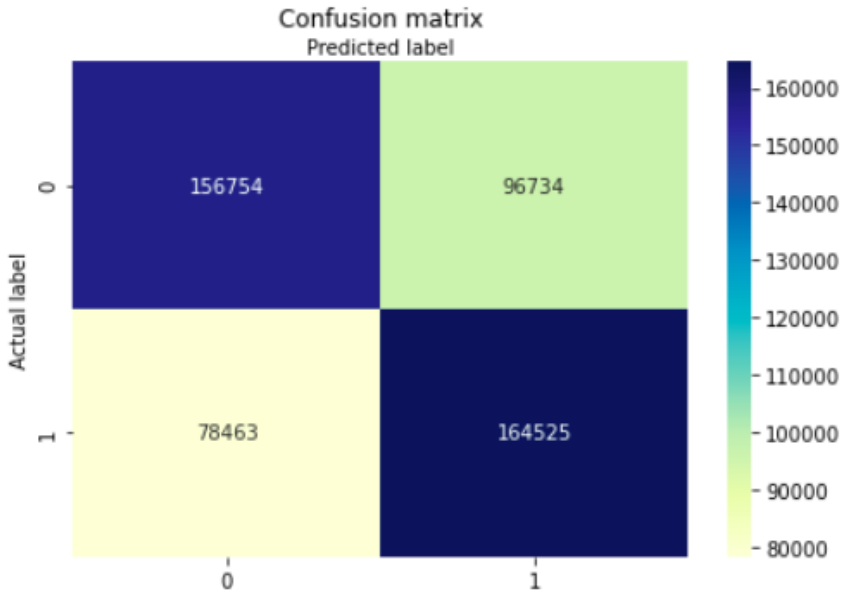
LZc



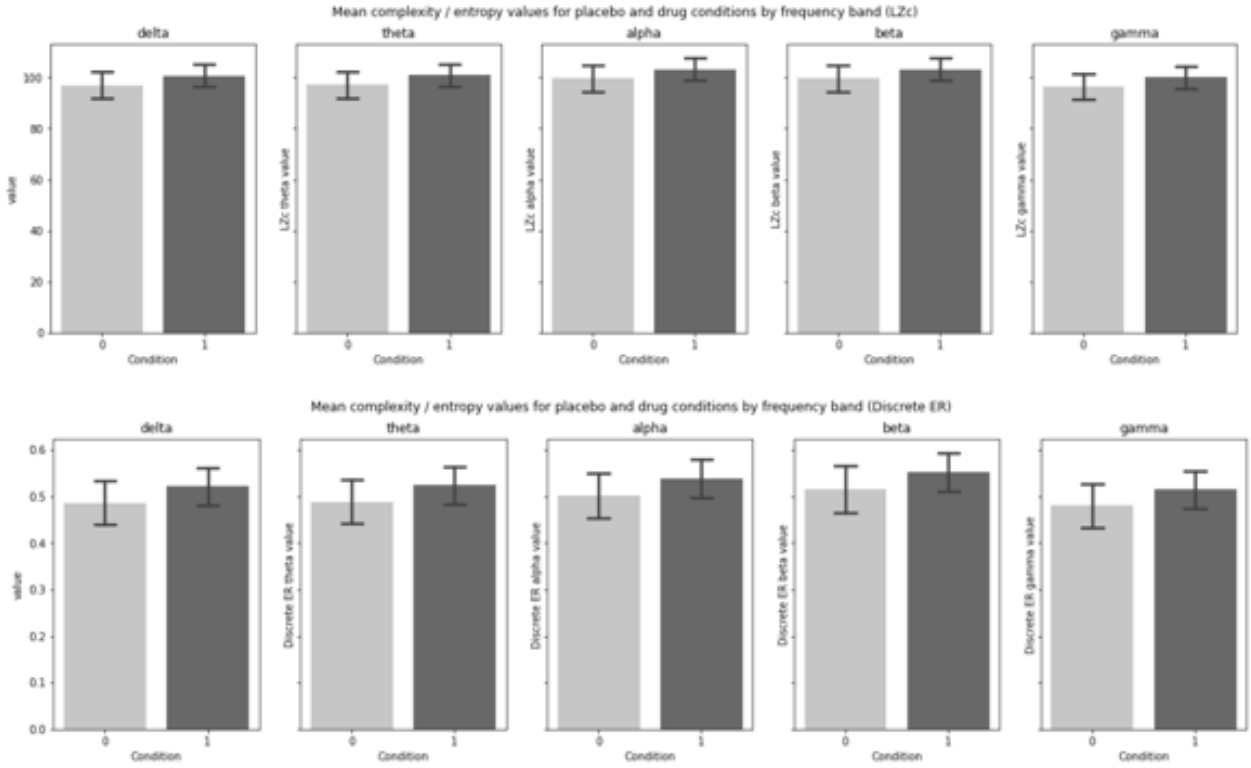
Discrete ER



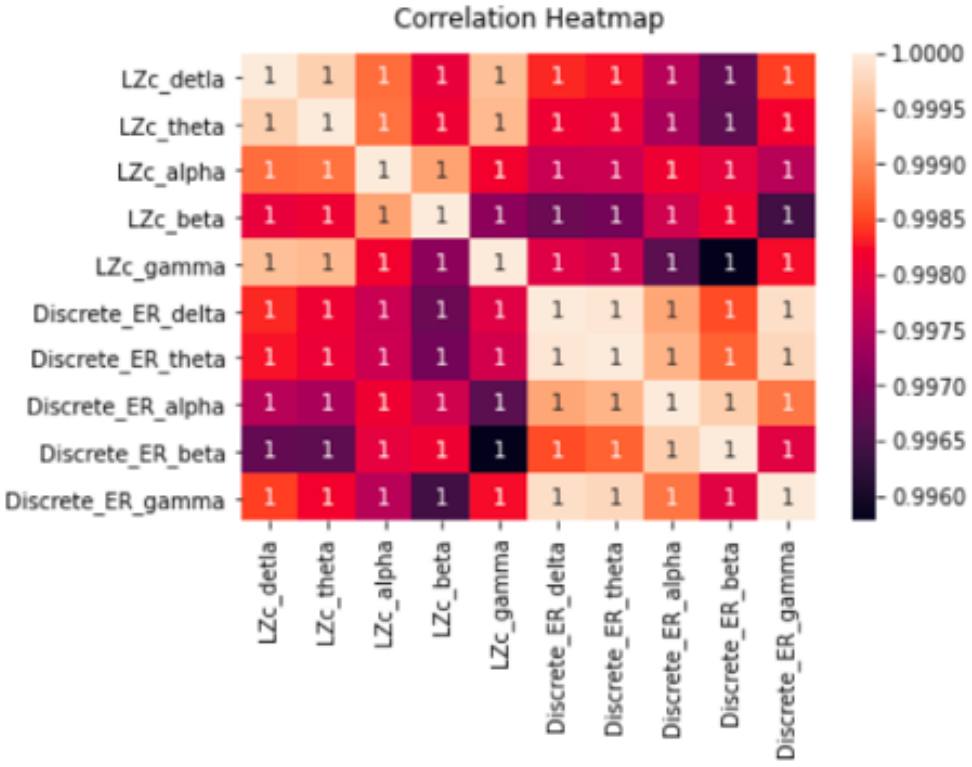
Appendix 4. Confusion matrices for models trained using CSER, LZc and Discrete entropy rate data. Misclassifications are the sum of the bottom left box and top right box. Correct classifications are the sum of the bottom right and top left box.



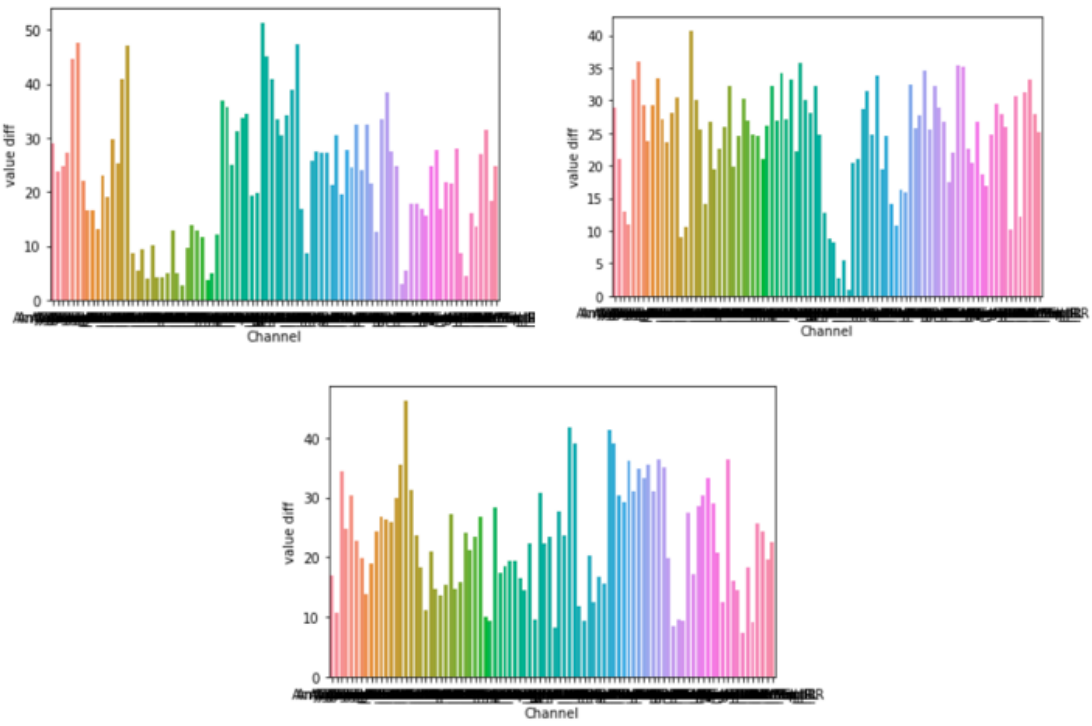
Appendix 5. Confusion matrix for model after outlier removal



Appendix 6. Bar graphs showing no complexity difference between conditions for the five frequency bands for the LZc (top) and Discrete entropy rate (bottom) measure

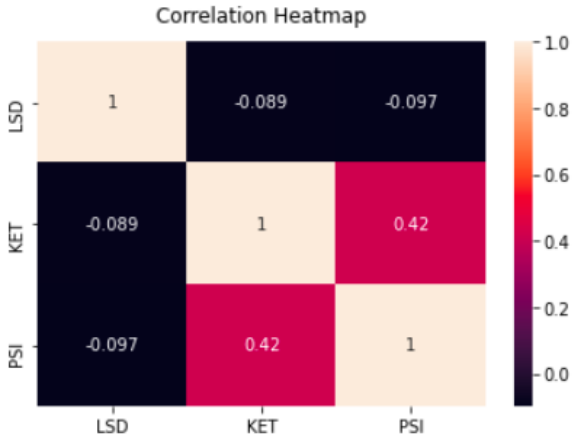


Appendix 7. Correlation heatmap for the 5 frequency bands and 2 measures

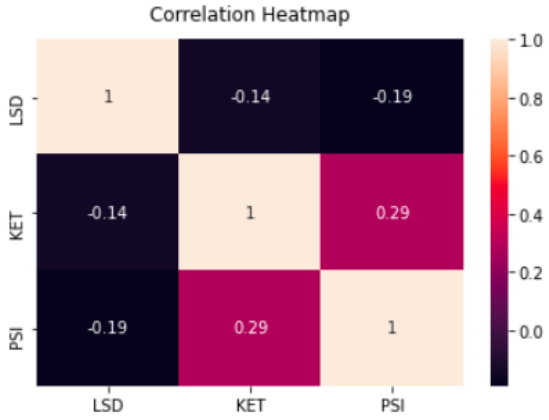


Appendix 8. The value difference in complexity between conditions for each channel for each drug type. Top left is PSI. Top right is LSD and bottom is KET.

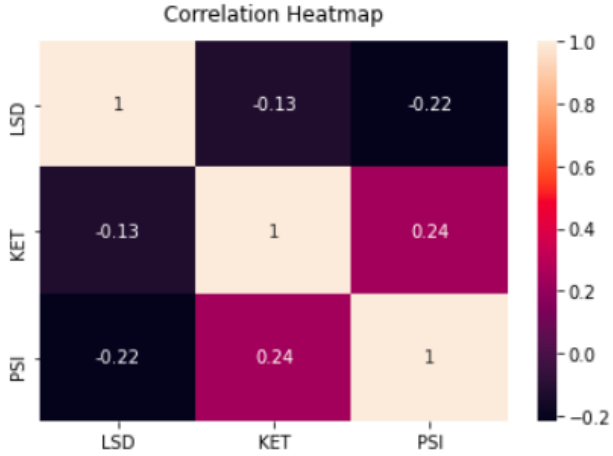
CSER



LZc



Discrete ER



Appendix 9. Correlations between the three drug types for each measure.

CSER

	Original 300Hz model	150Hz	200Hz	600Hz
Accuracy	.647	.634	.646	.599
Precision	.630	.616	.629	.584
Recall	.677	.649	.678	.635
F1 Score	.647	.634	.646	.599
AUC-ROC	.648	.635	.647	.600

LZc

	Original 300Hz model	150Hz	200Hz	600Hz
Accuracy	.628	.609	.628	.626
Precision	.608	.581	.608	.612
Recall	.676	.727	.676	.649
F1 Score	.628	.604	.628	.626
AUC-ROC	.629	.611	.629	.627

Discrete ER

	Original 300Hz model	150Hz	200Hz	600Hz
Accuracy	.631	.623	.627	.625
Precision	.613	.602	.605	.609
Recall	.664	.672	.670	.650
F1 Score	.631	.622	.626	.625
AUC-ROC	.632	.624	.628	.625

Appendix 10. Test metrics for models trained with varying sampling frequencies