School of Electrical and Mechanical Engineering

Department of Mechanical, Energy, and Industrial Engineering

Student email address: kaloso.tlotleng@studentmail.biust.ac.bw

CORRELATING SUBSTANCE USE DISORDER CLASSIFICATIONS FROM CLINICAL ASSESSMENTS AND BRAIN-COMPUTER INTERFACE EXPERIMENTS USING UNSUPERVISED MACHINE LEARNING

by

KALOSO MPHO TLOTLENG

Student ID number: 14001854

BEng (Mechatronics) (BIUST)

A Dissertation/Thesis Submitted to the School of Electrical and Mechanical Engineering in Partial Fulfilment of the Requirements for the Award of the Degree of Master of Engineering in Mechatronics and Industrial Instrumentation of BIUST

Supervisor(s): Professor Rodrigo S. Jamisola Jr.

Department of Mechanical, Energy, and Industrial Engineering

School of Electrical and Mechanical Engineering, BIUST

E-mail Address: jamisolar@biust.ac.bw

April, 2025

BOTSWANA INTERNATIONAL UNIVERSITY OF SCIENCE & TECHNOLOGY

POSTGRADUATE SCHOOL

Declaration Regarding the Work and Copyright

DECLARATION REGARDING THE WORK AND COPYRIGHT

Candidate (please write in caps or type) _____KALOSO MPHO TLOTLENG _____

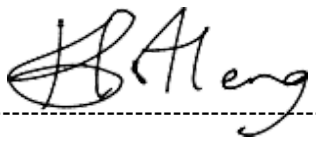
Student ID: _____14001854_____

Thesis Titled: _____CORRELATING SUBSTANCE USE DISORDER CLASSIFICATIONS FROM CLINICAL ASSESSMENTS AND BRAIN COMPUTER INTERFACE EXPERIMENTS USING UNSUPERVISED MACHINE LEARNING_____

I, the **Candidate**, certify that the thesis is all my original work and that I have not obtained a degree from this university or elsewhere based on any of this work.

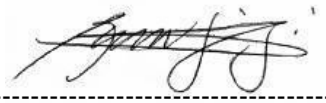
(If the thesis is based on a group project, then the student must indicate the extent of her / his contribution, with reference to any other theses submitted or published by each collaborator in the project, and a declaration to this effect must be included in the thesis)

This dissertation/thesis is copyright material protected under the Berne Convention, the Copyright and Neighbouring Rights Act, Act. No. 8 of 2000 and other international and national enactments, on that behalf, on intellectual property. It must not be reproduced by any means, in full or in part, except for short extracts in fair dealing; for researcher private study, critical scholarly review, or discourse with an acknowledgment, without the written permission of the office of the Postgraduate School, on behalf of both the author and the BIUST.

Signed:  Date: ----------23/04/2025----------

Primary Supervisor (please write in caps or type) _____RODRIGO S. JAMISOLA JR._____

I, the Candidate's **Primary Supervisor**, hereby confirm that I have inspected the above-titled thesis and, to the best of my knowledge, it is based on the original work of the candidate.

Signed:  Date: ----------24/04/2025----------

Acknowledgement

I would like to give a special thank you to my family for their unwavering support during my Master's degree journey. I am mostly grateful to my supervisor Prof. Rodrigo S. Jamisola Jr. for his dedication in mentoring me to develop new ideas using machine learning that address some of the societal problems faced in the world. His relentless motivation, guidance, and support in achieving our set objectives was impeccable. In addition, I would also like to appreciate the support from Prof. Jennifer L. Brown and Prof. Berdakh Abibullaev for their collaborative efforts in supporting my research study. I want to acknowledge the Ministry of Health and Wellness and the BIUST Human Research Ethics Committee for assisting me with the required permits to conduct human experiments. Not to forget Ms. Keneilwe Ogopotse and Mr. Mothusi Mashumba for their full cooperation in ensuring that my requests for assistance from their staff were well received and supported. Furthermore, I am grateful to the BIUST Office of Research, Development, and Innovation for their funding support under project code S00304 which has allowed me to have the necessary equipment to develop my research study. Lastly, I would like to thank all those who volunteered to be included in my brain-computer interface experiments using alcohol stimulus. Due to the Ethics of Human Research, the identity and information of these participants have been kept confidential.

Publications

The majority of this thesis comprises the research that has been submitted and is under review in international journals. All the research objectives, methodology, literature review, and results are presented in the thesis as they have been in the journal papers.

A. **Journal Papers**

1. Tlotleng K.M and Jamisola R.S, *An Analysis of the Severity of Alcohol Use Disorder Based on Electroencephalography using Unsupervised Machine Learning*, Big Data and Cognitive Computing, (**Submitted and under Review**).

2. Tlotleng K.M, Jamisola R.S, Brown J.L, *Correlating Clinical Assessments for Substance Use Disorder using Unsupervised Machine Learning*, Behavioral Sciences, (**Submitted and under Review**).

B. **Conference Papers**

1. Tlotleng K.M and Jamisola R.S (2023), *An Interpretation of the Transitions in EEG Signals Based on the Five Frequency Bands with Increasing Alcohol Content in the Human Body*, BIUST Teaching Research and Innovation Symposium (TRDAIS), (**Accepted**).

Contents

Declaration of Copyright i

Acknowledgement ii

Publications iii

List of Figures xiv

List of Tables xv

List of Abbreviations xvii

Abstract xviii

1 Introduction 1

1.1 Background . . . 2

1.1.1 Clinical assessments for SUD . . . 2

1.1.2 Brain-Computer Interface Experiments . . . 2

1.2 Problem Statement . . . 3

1.3 Aim and objectives . . . 4

1.3.1 **Aim** . . . 4

1.3.2 **Research Objectives** . . . 4

1.4 Hypothesis . . . 5

1.5 Significance of the Study . . . 5

1.6 Contribution to the body of knowledge . . . 6

1.7 General Outline Of the Thesis . . . 7

2 Literature Review 8

2.1 Clinical assessments for substance use disorders . . . 8

2.2 Brain-computer interface using alcohol stimulus . . . 13

3 Methodology 19

3.1 Overall methodology . . . 19

3.1.1 Outline . . . 19
3.1.2 Gather and analyze data of clinical assessments based on DSM V . 19
3.1.3 Perform human experiments given alcohol stimulus . . . 20
3.2 National Survey on Drug Use and Health (NSDUH) 2019 online database . 20
3.2.1 Data Preprocessing . . . 21
3.2.1.1 Data Cleaning . . . 22
3.2.1.2 Feature Correlation . . . 22
3.2.1.3 Feature Scaling . . . 24
3.2.1.4 Dimensionality Reduction . . . 24
3.2.1.5 Data-matching . . . 25
3.2.2 Cluster Analysis . . . 25
3.2.2.1 K-means Clustering . . . 26
3.2.2.2 Hierarchical Clustering . . . 26
3.2.2.3 DBSCAN . . . 27
3.2.3 Validation . . . 28
3.2.3.1 Internal validation . . . 28
3.2.3.2 External Validation . . . 28
3.3 Brain-computer interface experiments using alcohol stimulus . . . 29
3.3.1 Data gathering with participants . . . 29
3.3.1.1 Demography of participants . . . 29
3.3.1.2 Eligibility criteria . . . 30
3.3.1.3 Self assessment . . . 30
3.3.1.4 Experimental Device . . . 31
3.3.1.5 Experiment procedure . . . 31
3.3.2 Data preprocessing . . . 32
3.3.2.1 Importing raw data . . . 32
3.3.2.2 Filtering . . . 33
3.3.2.3 Epoching . . . 33
3.3.2.4 ICA . . . 33
3.3.2.5 Discrete wavelet transform . . . 34
3.3.3 Dimensionality reduction . . . 34
3.3.3.1 Principal Component Analysis . . . 34

3.3.4 Clustering . . . 34
3.3.4.1 BIRCH . . . 34

4 Results and Discussions **36**
4.1 Clinical assessments for SUD . . . 36
4.1.1 Alcohol dataset . . . 36
4.1.1.1 K-means clustering . . . 36
4.1.1.2 Hierarchical clustering . . . 38
4.1.1.3 DBSCAN . . . 39
4.1.2 Marijuana dataset . . . 40
4.1.2.1 K-means clustering . . . 41
4.1.2.2 Hierarchical clustering . . . 42
4.1.2.3 DBSCAN . . . 42
4.1.3 Cocaine dataset . . . 43
4.1.3.1 K-means clustering . . . 43
4.1.3.2 Hierarchical clustering . . . 44
4.1.3.3 DBSCAN . . . 45
4.1.4 Validation Results . . . 45
4.2 Interpretation of Results . . . 48
4.2.1 By Machine Learning . . . 48
4.2.2 By a Mental Health Professional . . . 53
4.3 Brain-computer interface experiments using alcohol stimulus . . . 53
4.3.1 Power spectral density analysis . . . 54
4.3.1.1 Theta band . . . 54
4.3.1.2 Alpha band . . . 57
4.3.1.3 Beta band . . . 60
4.3.2 Transition based clustering results . . . 65
4.3.3 Overall clustering results . . . 68

5 Conclusion and Future Directions **71**

A Feature Set – Alcohol Dataset **73**

Alcohol Dataset Features **74**

Alcohol Dataset Features **75**

B Feature Set – Marijuana Dataset **76**

Marijuana Dataset Features **77**

Marijuana Dataset Features **78**

C Feature Set – Cocaine Dataset **79**

Cocaine Dataset Features **80**

Cocaine Dataset Features **81**

D Power spectral density graphs for all the 17 participants **82**

E Approval Letter from Ministry of Health and Wellness **83**

Initial Research Permit **84**

Initial Research Permit **85**

Final Research Permit **86**

F Internal Approval Letter from the office of the Vice Chancellor **87**

Internal Permit **88**

G Ethical Clearance Certificate from Human Ethics Research Committee **89**

Ethical Clearance Certificate **90**

H Approval from the BIUST Clinic under the Division of Student Affairs and Campus Life **91**

Approval for Nurse to Assist in Human Experiments **92**

I Approval from the Directorate on Physical Planning and Campus Facilities **93**

Approval for Two Male Security Officers to Assist in Human Experiments **94**

J Approved Consent form for the Alcohol drinking experiments 95

Approved Consent Form 96

Approved Consent Form 97

Approved Consent Form 98

Approved Consent Form 99

K The Alcohol Use Disorders Identification Test used in the experiments to test the frequency and rate of use of alcohol for the participants 100

Alcohol Use Disorders Identification Test 101

List of Figures

3.1 The framework showing the proposed model structure to analyze data from an online database. The arrows indicate the sequence of our approach where preprocessing, cluster analysis, and validation algorithms were applied consecutively for each dataset. There are three substances considered, namely, alcohol, marijuana, and cocaine. Three unsupervised machine learning algorithms are used, namely, k-means clustering, hierarchical clustering, and DBSCAN. 21

3.2 The figure shows the correlation matrix for the alcohol dataset. A more dominant and positive correlation can be observed from the 20th feature to the last feature which is depicted by the concentrated purple color. 23

3.3 The graph shows the correlation matrix for the marijuana dataset. The graph is almost similar to the alcohol, except the features positively correlated can be observed from the 18th to the last feature. 23

3.4 The figure shows the correlation matrix for the cocaine dataset. This graph also has similar characteristics to the alcohol and marijuana correlation matrices in the sense that the highly correlated features are those under the substance abuse and dependence section. 24

3.5 The block diagram showing the different steps taken to analyze the raw EEG data. The arrows indicate the sequence of our proposed framework, starting with the recording of brain signals, followed by processing, and results analysis. We use the BIRCH algorithm to make a classification for transition-based and overall clustering results. 29

3.6 The Emotiv EPOC Flex Gel Sensor 32-channel EEG headset used to harness the brain activity of participants during the BCI experiments with alcohol stimulus 31

3.7 A figure showing the montage standard electrode placement for a 32-channel EEG headset that follows the 10-20 system 32

3.8 The raw EEG signals as recorded from the brain of a subject during the alcoholism experiments. It shows the 32 channels and their activity in continuous time 33

4.1 The figure represents the elbow method under the alcohol dataset to determine the optimal number of clusters. The graph shows a bend at k=2, indicating two clusters in the dataset. A smooth line immediately after the inflection point can be observed, which is an indication that it is where our model fits best. 37

4.2 The k-means clustering for the alcohol dataset showing two distinct groups as per the elbow method. Some data points are closer to the boundary across the groups but have been assigned to their respective clusters. It indicates that the model chose the lesser Euclidean distance for each point. Overall, the red cluster has more points to show the majority of subjects belong to one group. 37

4.3 The dendrogram for the optimal number of clusters under hierarchical clustering is shown in the figure. The structure of the hierarchy shows an approximately uniform distribution with two clusters, however, we notice that the first group has more congregated data than the second one. 38

4.4 The hierarchical clustering for the alcohol dataset is shown in the figure. Results of the clustering show that the first group has more data points than the second and the model confirms the two clusters analogy by the dendrogram. 38

4.5 DBSCAN results for the alcohol dataset showing three distinct clusters with noise. The structure of the data points distribution has not changed for all three algorithms. Our analysis indicates that we have three levels of AUD for this algorithm. 39

4.6 The elbow method under the marijuana dataset with a smooth line after the elbow of the graph to define two clusters for visualization. 40

4.7 The k-means cluster results for the marijuana dataset are shown. The few points between the clusters have greater Euclidean distance from the centroid but still form part of the respective clusters. More respondents are belonging to the red cluster than the blue cluster. 40

4.8 The dendrogram shows the density of points between each cluster. Based on the graph, we have two unique visualizations and the green cluster is more widely spread than the orange cluster. 41
4.9 Hierarchical clustering results for the marijuana dataset are shown. Unlike for k-means, the two clusters are not bounded around 0 and the data points within their boundary have been assigned to the orange cluster in this graph. 41
4.10 DBSCAN for the marijuana dataset is shown in the figure. There is no noise cluster meaning that we have no outliers in the dataset. There are many different groups, and the majority of data points belong to cluster 2. 42
4.11 The elbow method for the optimal number of clusters which is almost similar to the previous datasets. We obtain a very smooth line immediately after k=2 to indicate two clusters for visualization. 43
4.12 The k-means clustering results for the cocaine dataset are shown with well-spaced clusters. The intra-cluster distances however are more spaced in the green cluster. 43
4.13 The dendrogram shows a major difference between the two clusters as the majority of data points belong to the first cluster. The figure also indicates that points in the second cluster should be closely distributed. 44
4.14 Hierarchical clustering for the cocaine dataset with a high density of points in the first cluster. Although we expected a close relationship between the data points, in the figure we can observe a separation between some points with one not forming part of the subgroups. This structure is similar to that of k-means in Fig. 4.12. 44
4.15 DBSCAN for the cocaine dataset with many clusters assigned very close to each other. 45
4.16 Epoch 1 represents the initial phase of alcohol intake for the low-risk consumption in the theta band. We can observe a range of 0.065 μV²/Hz PSD and a high degree of signal entanglement. 55
4.17 Epoch 6 represents the last phase of alcohol intake for the low-risk consumption in the theta band. The average PSD range was reduced by 0.01 μV²/Hz. The degree of signal entanglement is less diverse compared to epoch 1. . . 55

4.18 Epoch 1 represents the initial phase of alcohol intake for the harmful or hazardous consumption in the theta band. We can observe a range of 0.08 μV^2/Hz PSD and a lesser degree of signal entanglement compared to the low-risk consumption. 56
4.19 Epoch 6 represents the last phase of alcohol intake for the harmful or hazardous consumption in the theta band. The average PSD range increased by 0.065 μV^2/Hz. The degree of signal entanglement is less diverse compared to epoch 1. 56
4.20 Epoch 1 represents the initial phase of alcohol intake for the likelihood of AUD in the theta band. We can observe a range of 0.10 μV^2/Hz PSD and the degree of signal entanglement is less and the lower set of signals shows more stability. 57
4.21 Epoch 6 represents the last phase of alcohol intake for the likelihood of AUD in the theta band. The average PSD range increased by 0.14 μV^2/Hz between epochs 1 and 6. The degree of signal entanglement is less diverse in both epochs compared to the low-risk and harmful consumption. 57
4.22 Epoch 1 represents the initial phase of alcohol intake for the low-risk consumption in the alpha band. We can observe a range of 0.014 μV^2/Hz PSD and a higher degree of signal entanglement. 59
4.23 Epoch 6 represents the last phase of alcohol intake for the low-risk consumption in the alpha band. The average PSD range was increased by 0.002 μV^2/Hz. The degree of neural overlap is almost similar to epoch 1. 59
4.24 Epoch 1 represents the initial phase of alcohol intake for the harmful or hazardous consumption in the alpha band. We can observe a range of 0.024 μV^2/Hz PSD and a lesser degree of signal entanglement compared to the low-risk consumption. 60
4.25 Epoch 6 represents the last phase of alcohol intake for the harmful or hazardous consumption in the alpha band. The average PSD range increased by 0.024 μV^2/Hz between epochs 1 and 6. The degree of signal entanglement is more compared to epoch 1. 60

4.26 Epoch 1 representing the initial phase of alcohol intake for the likelihood of AUD in the alpha band. We can observe a range of 0.023 μV^2/Hz PSD and a lesser degree of signal entanglement compared to the low-risk and harmful consumption. The lower set of signals have a smooth and stable trend while the higher signals are more jumpy. 61
4.27 Epoch 6 representing the last phase of alcohol intake for the likelihood of AUD in the theta band. The average PSD range increased by 0.053 μV^2/Hz between epochs 1 and 6. The degree of neural entanglement is less diverse in both epochs compared to the low-risk and harmful consumption. 61
4.28 Epoch 1 represents the initial phase of alcohol for the low-risk consumption in the beta band. We can observe a range of 0.0095 μV^2/Hz and a higher degree of spiky signals. 63
4.29 Epoch 6 represents the last phase of alcohol intake for the low-risk consumption in the beta band. The average PSD increased by 0.035 μV^2/Hz with a there is a significant spike at 15Hz. 63
4.30 Epoch 1 represents the initial phase of alcohol for the harmful or hazardous risk consumption in the beta band. We can observe a range of 0.065 μV^2/Hz and a lower degree of signal spikes compared to the low-risk consumption. We can observe a decrease in the PSD of these signals as the frequency increases. 64
4.31 Epoch 6 represents the last phase of alcohol intake for the harmful or hazardous risk consumption in the beta band. The average PSD increased by 0.0135 μV^2/Hz with a lesser degree of signal spikes. Similar to epoch 1, a decrease in the PSD of these signals as the frequency increases is observed. 64
4.32 Epoch 1 represents the initial phase of alcohol intake for the likelihood of AUD in the beta band. We can observe a PSD of 0.013 μV^2/Hz and more stability in the lower set of signals. The signals spikes are less significant and the value of the PSD increases as the frequency decreases. 65

4.33 Epoch 6 represents the last phase of alcohol intake for the likelihood of AUD in the beta band. The average PSD increased by 0.011 $\mu V^2/Hz$ and there is more signal fluctuations compared to epoch 1. Similar to the harmful or hazardous risk consumption, the power spectrum reduces as the frequency increases. 65
4.34 The clustering results in the first instance of alcohol consumption for all 17 participants. The green cluster has more dispersed data points compared to the red and blue clusters. 67
4.35 The clustering results in the last instance of alcohol consumption for all 17 participants. In this graph, the data points are more confined and have a lesser dispersion. 67
4.36 The overall clustering results of all the 17 participants from the first phase to the last phase of alcohol intake. The three clusters represent the brain activity of people with hazardous alcohol use, dependence symptoms, and lastly severe alcohol use. 69

List of Tables

2.1 Summary of machine learning studies related to substance use disorder, detailing data types, mental health focus, ML models, and evaluation metrics . 17

2.2 Summary of EEG-based studies on alcohol use disorder detection using machine learning, including experimental setup, signal features, and classification performance 18

4.1 Internal and External validation scores: Classical algorithms 46

4.2 Internal and External validation: State-of-the-art algorithms 46

4.3 Interpretations of results by Machine Learning and by a Mental Health Professional. 49

4.4 Interpretations of results continued from Table 4.3 50

4.5 Observations and the Interpretations of the PSD results obtained from the theta frequency rhythm. 54

4.6 Observations and the Interpretations of the PSD results obtained from the alpha frequency rhythm. 58

4.7 Observations and the Interpretations of the PSD results obtained from the beta frequency rhythm. 62

4.8 Observations and the Interpretations of the transition-based clustering results. 66

4.9 The three domains of the AUDIT with the total responses recorded from the 17 participants in the alcoholism experiments. 68

List of Abbreviations

AA	Alcohol Abuse
AD	Alcohol Dependence
ANN	Artificial Neural Network
AUD	Alcohol Use Disorder
AUDIT	Alcohol Use Disorders Identification Test
ASD	Autism Spectrum Disorder
BCI	Brain-Computer Interface
BIRCH	Balanced Iterative Reducing and Clustering using Hierarchies
CAD	Computer-Aided Diagnosis
CaUD	Cannabis Use Disorder
CD	Correlation Dimension
CoUD	Cocaine Use Disorder
CSV	Comma Separated Variables
DBSCAN	Density-Based Spatial Clustering of Applications with Noise
DSM	Diagnostic and Statistical Manual
DWT	Discrete Wavelet Transform
EDF	European Data Format
EC	Eyes Closed
EEG	Electroencephalography
EO	Eyes Open
FAWT	Flexible Analytical Wavelet Transform
FIF	Functional Imaging File
ICA	Independent Component Analysis
LDA	Linear Discriminant Analysis

LMT	Logistic Model Trees
LS-SVM	Least Square -Support Vector Machine
LR	Logistic Regression
MAPE	Mean Absolute Percentage Error
NESARC	National Epidemiologic Survey on Alcohol and Related Conditions
NSDUH	National Survey on Drug Use and Health
PCA	Principal Component Analysis
PSD	Power Spectral Density
RMSE	Root Mean Squared Error
SVM	Support Vector Machines
SUD	Substance Use Disorder
UCIKDD	University of California, Irvine Knowledge Discovery Database

Abstract

This work presents two investigations of clustering human behaviors resulting from substance use which are interpreted from graphs created using unsupervised machine learning. These investigations involve two datasets namely, the clinical assessments dataset and the brain-computer interface experiments dataset. The clinical assessments dataset is obtained from the National Survey on Drug Use and Health online database which is based on the Diagnostic and Statistical Manual version 5.0 (DSM-5) of mental health disorders. On the other hand, the brain-computer interface experiments dataset is obtained from an experiment of recording electroencephalogram (EEG) signals from volunteers while they consume alcoholic beverages for a specific period of time. The advantage of using unsupervised machine learning in this study is that it is capable of automatically searching for hidden patterns or groupings resulting in correlations within the dataset. These correlations are shown in the clustering graphs that visualize the relationships among human behaviors for analysis. In the clinical assessments dataset, we deduce the corresponding human behaviors from the data correlations shown in the clustering graphs. These interpretations are then validated by a mental health professional. In the brain-computer interface experiments dataset, the recorded EEG signals are analyzed using signal processing, and the corresponding human behaviors are deduced. These behaviors are then validated using the clustering graphs resulting from unsupervised machine learning.

1. Introduction

The rapid growth in SUDs has made it more difficult to contain and deal with the dangers they pose. The drugs ”hijack” the system and induce potent forms of reinforcement learning which results in aberrant and negative consequences [1, 2]. It is however noticeable that a lot of effort has been undertaken to help tackle the effects associated with SUD. The introduction of machine learning and standard tests to the field of psychiatry has immensely improved the efficiency of diagnosis as researchers can collect real-time data to identify target groups for interventions [3], thus making it easier to detect a person with or without SUD. However, the major limitation of these methods is that they have not fully explored the neurocognitive behavior of people with SUD. This study proposes an EEG-based method for correlating SUD status and a person's brain signal activity to further understand behavioral patterns associated with each type of substance dependence based on DSM V criteria for SUD. This approach will make use of a Brain-Computer interface to carry out experiments for harnessing EEG signals to further understand the AUD status of people who drink alcohol.

1.1 Background

1.1.1 Clinical assessments for SUD

Substance use disorder (SUD) is a major concern to the health sector due to the many harmful effects it causes to the individual. The drugs induce reinforcement of negative behaviors that include drug-seeking and drug-taking [1]. Due to the serious consequences of these substances to the user, it was imperative to implement predictive measures that can help manage individual behavior. Machine learning is becoming a popular predictive tool in psychiatry given the clinical data related to SUDs [3]. Such that given new data, the user can be diagnosed from the machine learning model that was created from the existing data.

Many of the related studies have adopted means to diagnose whether a person has SUD or not based on some deterministic and predefined set of characteristics. The use of drugs and substances has led to subjects becoming reliant on them thus causing SUD, which immensely impacted brain activity and the resulting behavior. There are distinct differences in how SUD affects the body, and there exist correlating effects of different substances on particular human behaviors [4]. These are standardized by the Diagnostic and Statistical Manual (DSM) of mental disorders [5] that is used as a guide in analyzing clinical data for SUD.

1.1.2 Brain-Computer Interface Experiments

AUD is a type of SUD that is caused by consistent consumption and severe use of alcohol that can induce clinically significant impairment and distress patterns [6]. It remains the most chronic and highly prevalent disorder that contributes to adverse problems globally [7, 8]. Different drinking patterns exist, but those associated with AUD include binge drinking and heavy alcohol use for legal drinkers. Consumption of 5 or more alcoholic beverages for men and 4 or more for women in a short span at least once a month defines binge drinking [9]. Having 5 or more drinks on any day or 15 or more per week for men defines heavy alcohol use. Similarly, having 4 or more drinks on any day or 8 or more per week describes heavy alcohol use in women. Drinking patterns associated with heavy alcohol use include extreme binge and heavy drinking [10, 11]. Individuals may project different AUD severity levels namely, mild, moderate, and severe symptoms depending on the number of diagnostic criteria endorsed [12] as standardized by the DSM 5 criteria of of mental health

disorders. These are influenced by the user's inability to manage alcohol craving, tolerance, and withdrawal [13, 14]. The manifestation of these symptoms causes severe cognitive, behavioral, and psychological problems if early intervention is not administered [15].

Different conventional screening tools, such as standardized questionnaires and clinical assessments, have been proposed by mental health professionals to aid in diagnosing AUD. However, these self-test assessments are manual, subjective, and less accurate as there is a potential lack of honesty or memory impairment of the alcohol user in relation to their total consumption [16, 17]. Therefore many studies have adopted automated screening methods like EEG and machine learning to understand the brain activity of alcoholics. These tools have proven promising results in classifying a person's AUD status [18]. EEG is a non-invasive method of brain-computer interface (BCI) that measures the electrical activity of the brain. BCI enables an interaction between the brain and various machines by using devices that can collect and interpret the signals. To achieve this communication, the electrodes of an EEG brain wear are placed on the scalp of an individual to record the activity of the brain in real-time with high accuracy and safety [19, 20].

Alcohol evokes some significant cognitive and behavioral changes in the human brain and thus with the help of EEG, crucial information about a person's AUD status can be interpreted. These transitions can be analyzed and interpreted using machine learning and deep learning models [21].

1.2 Problem Statement

Clinical assessments on substance use disorder have been complied by mental health professionals for many years, in hundreds of thousands entries in the dataset. Machine learning methods applied to those datasets mostly used supervised machine learning, with the assumption that the model is fed with the correct answer. This approach is not feasible all the time. Direct measurements of brain signals as a result of alcoholic beverage stimulus generates hundreds of thousands entries in the dataset as well, and would require machine learning methods to analyze them.

Our study aims to address the problems stated above by proposing the following:

1. Use an unsupervised machine learning method to cluster the effects of substance use through correlations of features within the clinical assessment dataset. The advantage

of this method is that it does not require knowledge of the correct answer, instead the machine learning model automatically search hidden patterns or groupings within the dataset.

2. Perform experiments to record brain signals from persons as they consume alcoholic beverages at an increasing amount for a specified period of time. We first analyze the features of the recorded brain signals using signal processing methods, and then confirm the results by correlating these features using unsupervised machine learning.

1.3 Aim and objectives

1.3.1 Aim

The main aim of this research is to study substance use disorders by correlating clusters based on hidden patterns within a given dataset. This approach is performed using unsupervised machine learning. The two datasets used in this work are:

1. Clinical assessments for substance use disorder based on DSM 5
2. EEG experiments using alcoholic beverages stimuli to volunteers with increasing consumption at a given time

1.3.2 Research Objectives

1. Unsupervised machine learning in clinical assessments dataset
    - Collect and process dataset from the National Survey on Drug Use and Health based on the Diagnostic and Statistical Manual 5 of substance use disorder
    - Apply unsupervised machine learning to establish the correlation of features within the dataset
    - Correlations shown in the graphs are interpreted as human behavior as a result of substance use
    - Validation of the human behavior interpretations by a mental health professional
2. Unsupervised machine learning in EEG experiments dataset

- Record EEG signals during experiments with alcoholic beverage stimulus consumed by volunteers
- Apply signal processing to features of the dataset and analyze the corresponding human behavior as a result of alcohol use
- Apply unsupervised machine learning to establish the correlation of features within the dataset
- Correlations shown in the graphs are interpreted as human behavior as a result of alcohol use
- Validate the human behavior established from correlations versus human behavior analysis based on signal processing features

1.4 Hypothesis

Through the use of unsupervised machine learning, clusters or data patterns can be established that shows correlation of features within the dataset. The generated graphs through unsupervised machine learning can be analyzed and interpreted to deduce resulting human behaviors as a consequence of substance use. Two datasets will be used: clinical assessments dataset and brain-computer interface experiments dataset. These datasets will be used to investigate the proposed method.

1.5 Significance of the Study

This research aims to provide an understanding of the brain signal activity of individuals and their SUD status when given a stimulus. It will bridge a gap between the correlation of clinical assessments of SUD as well as the neurocognitive behavior of individuals exhibited when they are under the influence. This work uses alcohol as the addictive stimuli in the experiments because, at the time of the study, it was the only legal substance. This can lead to better intervention on SUDs by mental health professionals by providing:

1. Efficiency in the diagnosis of SUD among subjects.
2. Assisting mental health professionals with better diagnosis.

3. Early treatment.

4. Increased understanding of brain activities of a person with SUD.

1.6 Contribution to the body of knowledge

This research study contributes to the body of knowledge by the use of unsupervised machine learning to:

1. Analyze data from the NSDUH 2019 based on the DSM 5. To the best of our knowledge, this is the only work that processed the data from NSDUH 2019 and used unsupervised machine learning to model them.

    - Unsupervised machine learning is important in this kind of study because it allows the model to create its classification based on the data correlations, instead of assigning the target labels beforehand.
    - At the time of the study, NSDUH 2019 was a more recent database with 56,137 entries. DSM 5 is more current and is widely used by mental health professionals in classifying SUDs.
    - The result of this study is an automated classification in terms of SUD based on DSM 5 and is validated by a mental health professional.

2. Analyze AUD severity levels of people who drink alcohol, with data from BCI experiments using alcohol stimulus and AUDIT assessment scores. This study aims to identify the potential risk of current alcohol consumption and the transitions that occur as the amount of alcohol increases. To the best of our knowledge, this is the only work that processed EEG data from alcohol-drinking experiments and used unsupervised machine learning to model them.

The reason for choosing the unsupervised machine learning method in this study is because the data provided in the NSDUH database and the one harnessed from EEG alcohol-drinking experiments do not have target labels in terms of SUD or non-SUD. The unsupervised method allows us to make classifications based on data correlations without the need to predefine class labels before processing.

1.7 General Outline Of the Thesis

The general outline of this thesis will be structured as follows:

Chapter 2 - Literature Review

In this chapter, detailed review of the relevant studies that have adopted and applied machine learning and EEG methods to diagnose SUDs will be introduced.

Chapter 3 - Methodology

This chapter will show the proposed model structure and extensively describe the tools and outline the process used to derive the required output. It will provide a detailed description of the approach in correlating clinical assessments from an online database and brain-computer interface experiments using alcohol stimulus. The results will be clustered using unsupervised machine learning.

Chapter 4 - Results and Discussions In this chapter, the graphical visualizations of the results will be shown for:

- Clinical assessments for SUD
- Brain-computer interface experiments using alcohol stimulus

The different approaches used to analyze data will be shown, including the observations, and the interpretations of the results. A full discussion of these results is provided in this section.

Chapter 5 - Conclusions and Future Direction

This chapter presents the major conclusions drawn from the study, including the recommendations for further work.

2. Literature Review

In this chapter, the detailed literature review is discussed which is subdivided into two sections.

2.1 Clinical assessments for substance use disorders

In the following paragraphs, this work presents studies in general addiction using supervised machine learning, the algorithm used to model the status of addiction, and the corresponding dataset used. None of these studies used DSM criteria. These studies are summarized in the first part of Table 2.1. A study by [29] introduced the development of a decision support system for detecting drug abuse using artificial neural networks (ANN) in which the experimental results obtained a classification accuracy of 98.42% for correct diagnosis. Data were obtained from 184 cases of individuals with SUD who were assigned with a set of questions compiled by specialists. In this study, a simple mechanism to help an investigator and families identify individuals with SUDs was deployed which would provide adequate information on the symptoms those individuals experienced. This machine-learning algorithm was used to find out if a person meets the criteria for SUD or not. An examination of the ability of machine learning methodologies to predict nicotine dependence was adopted by [30] in which they developed a classifier for female smokers. The study was aimed at determining how accurate machine learning would be in classifying the water-pipe smoking behaviors of women. As mentioned in the study, the authors opted to use support vector machines and decision trees to predict the level of nicotine dependence due to the difficulty of deploying unsupervised learning for the study. To supplement research on using machine learning for addiction studies, [31] proposed a method to predict alcohol use disorder (AUD) from electronic health records which utilized supervised learning. The study developed an automated classification model using supervised learning algorithms to classify data between alcohol use disorder (AUD) positive and AUD negative. The performance was measured using various metrics to verify the results. In the next few paragraphs, a discussion is presented for studies in mental health disorders using unsupervised machine learning, the algorithms used to model the mental health disorders, and the corresponding datasets used. Only one of the

studies presented used DSM criteria. A summary of these studies is shown in the second part of Table 2.1. The study [22] established a method that utilized the application of various machine learning algorithms to identify the state of mental health in a target group. It modeled the subjects' answers to a questionnaire using unsupervised machine learning. It formed relevant clusters that were verified by mean opinion score. Classification models were labeled that predicted the mental health of individuals. The study adopted k-means, hierarchical, and k-medoids clustering techniques. Due to little practical guidance available to measure the prediction strength of unsupervised machine learning cluster models, [23] proposed an extension of the cross-validation procedure to evaluate the quality of the clustering algorithm and to predict cluster membership for new data points. The paper focuses on understanding, how good the model is or to what extent had the model correctly assigned new data points to a relevant cluster, and measures the prediction strength of the model. It used k-fold cross-validation to calculate the probability of occurrence of the labels in both the training and testing data. The study further utilizes root mean squared error (RMSE) and mean absolute percentage error (MAPE) to evaluate the prediction strength of the algorithm. Calinski-harabasz score was used to obtain the optimal number of clusters and the model was tested using k-means. It identified three clusters for the emotions dataset, four clusters for the chronic disease dataset, and five clusters for the yeast dataset. The research showed promising results with an overall RMSE of less than 0.075 and MAPE of less than 12.5% for the three datasets. Identification and analysis of behavioral phenotypes in autism spectrum disorder (ASD) were carried out by [24] in which an unsupervised machine learning model was applied to ASD subgroups and their taxonomic relationships. In this study, a Gaussian mixture model was implemented and revealed 16 subgroups from a sample of *n=1034* that was a subset of *N=2400* of children with ASD. This portion of the trial was used as retrospective treatment data and the responses were examined using linear regression to understand how each sub-unit responded to varying treatment intensity. Hierarchical agglomerative clustering was performed before fitting the data into the regression model and the results revealed two overlying behavioral phenotypes. Within these patterns, unique and significantly low profiles contained subgroups that differed in severity. The high variance shown by regression models indicated that clustering provides a basis for similarity which would be ideal for administering treatment based on these cluster memberships. Finally, this work presents, in the next few paragraphs, mental health disorders using supervised

machine learning and statistical methods, and the corresponding datasets. All of the studies presented used DSM-IV and -5 criteria. A summary of these studies is shown in Table **??**. According to [32], there are distinct cognitive and neurobiological attributes between types of substance abuse. It compared opiate and stimulant addictions even though a full understanding of their specific effects has been limited by high poly-substance dependence. This study is an attempt to identify behavioral markers that can classify heroin and amphetamine dependence by using machine learning. A total of 222 participants had to meet a certain criterion as per the DSM-IV criteria [33] for drug addiction. The majority of subjects were in protracted abstinence and were classified across four different measures which include psychiatric, neuro-cognitive, demography, and personality. A total of 54 predictors under these classes were used and machine learning revealed substance-specific multi-variate profiles which classified heroin dependence (HD) and amphetamine dependence (AD) with high accuracy. Psychopathy emerged as the only classifier common to both types of addiction. A two-fold purpose was carried out by [34]. The main goal was to assess the reliability, validity, and classification accuracy of the DSM-5 diagnostic criteria for gambling disorder (GD). The second goal was to compare DSM-5 and DSM-IV on the above-mentioned parameters. Data comprised of two groups of subjects with GD and without GD. The results of reliability and classification accuracy are better in DSM-5 than DSM-IV. In terms of comparison, the study showed more similarities than differences between DSM-IV and DSM-5. These conclusions were achieved by initially computing values of the three indices for DSM-5 and then secondly comparing the psychometric measures of the two diagnostic criteria. Lastly, DSM-5 shows more superiority over DSM-IV for reliability, validity, and classification accuracy after removing subjects without GD. A study by [35] proposed an item response theory analysis to measure the performance of AUD and the quantity-frequency of alcohol consumption based on DSM-5 criteria. The main objective of the study was to examine the efficiency in terms of the overall diagnostic utility and the threshold, discrimination, information value, and differential criterion functioning of the DSM-5 criteria for AUD. Of the different patterns of alcohol use, the research focused solely on 5+/4+ binge drinking which was assessed at different levels of frequency. The results showed in-variance criteria amongst the different subgroups while also confirming the support for reliability and validity of DSM-5 concerning AUD (which included the 5+/4+ binge drinking). In addition to the above studies, more recent studies that have also established the application of

machine learning to SUDs are discussed. According to [25] machine learning was used to identify health, psychiatric, psychological, and contextual features that are essential for the prediction of SUD. These behavioral markers were also utilized to predict individuals with a high risk of developing SUD. A total of 600 participants (494 male and 206 female) were recruited and administered a questionnaire across several assessment waves that were carried out at the ages of 10-12, 12-14, 16, 19, and 22. The study used a random forest classifier to examine the characteristics that are mostly associated with SUD and thirty of those variables strongly emerged as the strongest predictors for SUD. The results showed that in 10-12-year-old youths, the features predicted SUD with an average of 74% which increased to 86% for the youth at the age of 22 years. Compared to other machine learning algorithms, the random forest classifier had the best performance in predicting individuals at risk of developing SUD between the ages of 10-22. The examination of clinical and functional characteristics as predictors of AUD levels was carried out by [26]. The study establishes the extent of alcohol use, comorbidity, and impairment on the mild, moderate, and severe AUD levels. A total of 588 participants aged 18 and above recruited from 2016 to 2019 were assessed for DSM 5 AUD for current substance use. In this work, multinomial logistic regression was used to examine the association between potential validators, psychiatric disorders, functional impairment, and the three AUD levels. The results indicated that craving, binge-drinking frequency, problematic use, and harmful drinking had a more significant possibility of being associated with the three AUD levels compared to the reference group that had no AUD or any other substance use. It was reported that all the potential validators were also associated with severe effects of AUD more than the mild and moderate effects of AUD. A study by [28] proposed a method to assess the degree of the existing relationships between the older and newer generations of illicit and prescription drugs. A total of 36,309 adults from wave 3 of the National Epidemiologic Survey on Alcohol and Related Conditions (NESARC-III) dataset aged 18 and above were assessed to evaluate the occurrence of the comorbidity between DSM 5 CUD, DSM 5 SUD, and substance use. Multivariable logistic regression results indicated that persons with current DSM 5 CUD are more prone to using all examined drugs for a lifetime including a year of several newer class illicit and prescription stimulant-based substances. It was also found that the current DSM 5 CUD could be related to the different DSM 5 SUDs with high comorbidity existing between them. By using 2015-2019 NSDUH data, [27] proposed a method to evaluate

gaps in care for persons with AUD. The framework was developed on adults based on the following measures: their prevalence of AUD, the proportion of those that received health care in the past year, and were screened for alcohol use, those with AUD that received brief intervention for misusing alcohol, those that were informed of alcohol misuse treatment, and lastly those that received treatment. The results were classified according to alcohol severity and the cascades showed that most individuals with AUD received health care in the past year. It was also reported that a minority of individuals received the intervention, treatment referral, and health aid. In classifying the severity of AUD similar patterns were observed across the cascades of care and it was observed that people with AUD normally do not receive treatment even after the screening on alcohol use. From Table 2.1, two previous studies are closest to the study presented in this work. First, the work in [23] is similar to the methodology that we adopted. They used unsupervised machine learning with one clustering algorithm, that is, k-means clustering. In addition, they used principal component analysis (PCA) for dimensionality reduction, and for validation techniques, they used silhouette score, k-fold cross-validation, RMSE, and MAPE. Compared to our work, they did not study SUD and did not use the publicly available data from the NSDUH database. Lastly, they did not use DSM criteria. The second study that is closest to our work is [35]. It considered SUD using only one type, that is, AUD, and used DSM-5 criteria. However, it used a logistic model that is a supervised machine learning technique. In summary, the studies [1, 3, 5] discussed mental health addictions and machine learning, [29, 30, 31] presented the general applications of machine learning in addiction studies, [22, 23, 24] showed the applications of unsupervised learning in addiction studies, [32, 34, 35] analyzed studies that used DSM and machine learning, and lastly [25, 26, 28, 27] are more recent studies that established the application of machine learning to SUDs based on DSM criteria. In general, there are considerable studies in addiction using machine learning but using different types of addiction, different machine learning methods, and different datasets. Further studies in this area have to consider the different elements of the previous addiction studies to make a contribution. This is what our proposed method is trying to achieve.

2.2 Brain-computer interface using alcohol stimulus

In the following paragraphs, the relevant studies that use EEG to detect AUD are discussed, including the experimental platforms utilized, the feature extraction methods used, and the corresponding machine learning algorithms. A study by [17] proposed a machine learning model to classify healthy controls and alcohol-dependent individuals. In this study, the quantitative EEG (QEEG) methods were developed to analyze and discriminate between the healthy controls, alcohol abusers, and alcoholics. A total of 45 subjects including 15 healthy controls, 18 alcoholics, and 12 alcohol abusers were recruited to participate in the experiments. The recording of the brain signals was done at a resting EEG state with 10 minutes of eyes open (EO) and eyes closed (EC) conditions. QEEG features were fed to Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), Multilayer back-propagation Network (MLP), Logistic Model Trees (LMT), and 10-fold cross-validation models. LMT achieved the best accuracy of 96% and the model managed to classify AUD patients from healthy controls with an accuracy of over 90%. A detection of AUD using EEG signal features and flexible analytical wavelet transform (FAWT) was carried out by [16] in which a computer-aided diagnosis (CAD) method was proposed. In this work, data from the University of California, Irvine Knowledge Discovery Database (UCIKDD) comprising a total of 122 non-alcoholic and alcoholic EEG subjects was used. The study uses FAWT alongside machine learning models to detect alcoholism using several features namely: mean, standard deviation, kurtosis, skewness, and Shannon entropy. These statistical features are used as input to Least Square -Support Vector Machine (LS-SVM), Support Vector Machine (SVM), and Naïve Bayes learners. LS-SVM performed better than the other models with an accuracy of 99.17%. According to [37], an EEG-based functional connectivity measure was utilized to automatically detect AUD between alcoholics and healthy controls. The study adopts the resting-state EEG method in which 513 features were extracted from 19 scalp locations using synchronization likelihood (SL). Data was collected from a total of 60 participants (30 healthy controls and 30 alcoholics) of similar age. Support Vector Machine (SVM), Naïve Bayesian Classifier(NBC), and Logistic Regression (LR) were used and SVM had the best performance of the three with an accuracy of 98%. A method to classify alcoholic EEG signals using Wavelet Decomposition and machine learning models was proposed by [38]. In this study, relevant features were extracted namely, standard deviation, power value, mean, minimum and maximum values, the ratio of the absolute value,

and the absolute mean. These were used as input to SVM, Optimum-Path Forest (OPF), Multilayer Perceptron (MLP), Naive Bayes, and k-Nearest Neighbors (k-NN). Data for the study consisted of a total of 77 male alcoholics and 48 right-handed males with no history of alcohol misuse or any psychiatric or neurological disorders. The model achieved an accuracy of 99.87% in classifying alcoholic EEG signals. An EEG frequency band-based features model was proposed by [41] for automatic screening and analysis of alcoholic EEG signals. The study applies intrinsic mode functions (IMFs) and empirical mode decomposition to the brain signal to measure and extract mean absolute deviation, interquartile range, entropy, coefficient of variation, and neg-entropy features. The results indicated that based on the p-value, features from low-frequency brain rhythms are more statistically significant in analyzing alcoholic EEG signals. These were used as input to extreme machine learning (ELM) and least squares support vector machines for the classification of alcoholic and normal EEG signals. ELM obtained an accuracy of 97.92% compared to the other methods. A study by [42] utilizes the correlation dimension (CD) method to extract features that are used to cluster alcoholic EEG signals. Data was obtained from the UCIKDD comprising a total of 122 normal and alcoholic EEG patients. Several distance feature selection metrics namely, Chebyshev, cosine, correlation, and city block were used to discriminate between normal and alcoholic EEG signals for non-linear features. Random Forest, Modified Adaboost.RT, Radial Basis Functions (RBF), MLP, SVM, KNN, K-means, and NBC classifiers were applied to the selected features. The model obtained an accuracy of 98.99% when the correlation distance is combined with CD and Modified Adaboost.RT classifier. In addition to the above literature we present, in the following paragraph relevant study closest to the proposed methodology of our work. Firstly, the work discussed by [43] is similar to the methodology we adopted. In this paper, the authors used discrete wavelet transform (DWT) to decompose the EEG signals into time-frequency features called wavelets. Furthermore, they implemented independent component analysis (ICA) for blind source separation and noise reduction. Finally, the study adopts principal component analysis (PCA) to perform dimensionality reduction. Compared to our work, this paper uses a publicly available dataset from the UCIKDD database. In our study, we harnessed data from the BCI experiments using alcohol stimulus recorded in real-time while the participant drinks alcohol. Secondly, their work uses supervised machine learning methods that include, KNN, LVQ, LSTM, B_LSTM, and CNN to classify between alcoholics and non-alcoholic individuals.

Our work adopts Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) algorithm to cluster the AUD status of our 17 participants based on their frequency and rate of alcohol consumption. An analysis of the results from brain activity using the power spectrum transitions from the first to last phase of alcohol consumption, and lastly, the overall clustering is provided. The relevant studies in Table 2.2 have proven to detect AUD using EEG signals and machine learning with high accuracy. Based on the results obtained, these are significant works that adopted numerous machine-learning algorithms, various feature extraction techniques, different experimental methods, and EEG datasets to make a classification between alcoholics and healthy controls. The studies [36, 17, 37] are closer to this work in that they have used EEG spectral power results to analyze and detect AUD from a group of alcoholics and non-alcoholics. The authors [36] claim that from a group of individuals with AUD, IGD, and healthy controls, more delta power was observed from those with AUD. There was, however, no correlation between results in theta power and the scores from the AUDIT-K assessment. In this work, there is a claim that the delta power can be used as a trait marker rather than a state marker. To compare individuals with AUD to healthy controls, [17] found a significant decrease of the AP in the brain of those with AUD. This reduction was found in the theta, alpha, gamma, and high gamma frequency bands. Additionally, the authors discovered that between alcoholics and alcohol abusers, the AP decreased significantly in the left occipital lobe. This study claims that the AP is better at discriminating between alcohol abusers, alcoholics, and healthy controls. Lastly, research by [37] states that spectral power can be an important tool to use in discriminating AUD patients with healthy controls. The authors further claim that results in the theta, beta, and high gamma are mostly significant when making a classification between healthy controls and individuals having AUD. Compared to our results, these studies do not record the brain activity of participants while they drink alcohol. They take a sample of alcohol and non-alcohol users and record their brain activity to make a classification between the two types of groups. This work analyzes the severity of AUD of people who are alcohol users and record their EEG as they drink alcohol. In studies [38, 16, 39], and [40], the UCI KDD EEG dataset is analyzed using different feature extraction techniques and machine learning models to detect alcoholism and AUD. Compared to our work, these studies do not record the brain activity of people while they drink alcohol, Additionally, the methods adopted in their research are also used to make a classification between people with and without AUD

similar to the studies discussed above. Our work is the only study that, uses data from a group of alcohol and records their brain signals as they drink alcohol.

In summary, the studies [16, 17], [37, 38], and [41, 42] have demonstrated the possibility of detecting AUD using EEG signals and machine learning with high accuracy. Based on the results obtained, these are significant studies that adopted numerous machine-learning algorithms, various feature extraction techniques, and different experimental methods to make a classification between alcoholics and non-alcoholics. Further studies in this work have to consider the different methods that are independent of a control/normal group to make substantive contributions to AUD diagnosis. This is what our proposed work is trying to accomplish.

In conclusion, existing research in SUD studies reveals several significant gaps. While many studies apply supervised machine learning techniques to address this problem, there is limited application of unsupervised machine learning methods that can reveal distinct data patterns without pre-assigned labels. Moreover, prior work has typically relied on conventional methods that separated clinical assessments, such as those based on DSM-5 and datasets like NSDUH, from EEG studies with alcohol use, and with EEG recordings generally obtained from pre-classified groups rather than during alcohol consumption. As a result, a more comprehensive approach that integrates clinical assessment data with real-time brain activity recording is lacking. This thesis aims to bridge these gaps by leveraging clustering techniques to analyze both clinical and experimental EEG data, thereby offering a novel, integrated method for the dynamic assessment of substance use disorder severity.

Title (First Three Words) Author and Year	Mental Health Issue	ML Models Used	Dataset	Features Extracted	Feature Selection	Final Features	Accuracy	Sensitivity	Specificity	F1 Score	Citations
Behavioral Modeling Mental *Srividya et al. (2018) [22]*	Mental illness risk classification (binary: risk/no-risk)	SVM, Naïve Bayes, KNN, Logistic Regression	Survey responses from 300 and 356 subjects	Demographic + behavioral survey data	Not specified	Not specified	90%	–	–	0.93	267
Cross-Validation Approach *Tarekegn et al. (2018) [23]*	Not specified	K-Means, RMSE, MAPE, Cross-Validation	Datasets on chronic illness, emotions, yeast	Not specified	Not specified	Not specified	–, –	–	–	–	28
Identification Analysis Behaviora *Stevens et al. (2020) [24]*	Autism Spectrum Disorder	Hierarchical Clustering	2400 U.S. children from community centers	Behavioral and assessment data	Not specified	Not specified	–	–	–	–	120
Analysis of Substance use *Jing et al. (2020) [25]*	Substance Use Disorder (SUD)	Random Forest	Recruitment by ads, calls, posters; age 10–22	Lifestyle, psychological, behavioral data	Not specified	Not specified	74–86%	–	–	–	60
Correlates of Mild, Moderate *Mannes et al. (2021) [26]*	Alcohol Use Disorder (AUD)	Multinomial Logistic Regression	588 participants from 2 medical centers	Demographic + clinical data	Not specified	Not specified	–	–	–	–	21
Cascade of Care *Mintz et al. (2021) [27]*	AUD	Multinomial Logistic Regression	Same as Mannes et al.	Demographic + clinical data	Not specified	Not specified	–	–	–	–	67
DSM-5 Cannabis Use *Hayley et al. (2017) [28]*	Cannabis Use Disorder (CaUD)	Logistic Regression	NESARC-III, n=36,309, ages 18+	Substance use patterns + demographics	Not specified	Not specified	–	–	–	–	67

[1] ANN – Artificial Neural Network,
[2] DNN – Deep Neural Network
[3] KNN – K-Nearest Neighbors
[4] MAPE – Mean Absolute Percentage Error
[5] NESARC – National Epidemiologic Survey on Alcohol and Related Conditions
[6] RMSE – Root Mean Squared Error
[7] SVM – Support Vector Machine

Table 2.1: Summary of machine learning studies related to substance use disorder, detailing data types, mental health focus, ML models, and evaluation metrics

Title (First Three Words) Author and Year	Mental Health Focus	EEG Experiment Type	ML Models Used	Dataset Description	Features Extracted	Feature Selection	Accuracy	Sensitivity	Specificity	F1 Score	Citations
Neurophysiological Features *Son et al. (2015) [36]*	IGD and AUD classification	Resting-state EEG; no task	GEE, Naive Bayes, KNN, Logistic Regression	76 subjects: 34 IGD, 17 AUD, 25 control	Frontal alpha power, beta coherence, etc.	Not specified	–	–	–	–	89
Automatic diagnosis of alcohol *Mumtaz et al. (2016) [17]*	AUD (multi-class: abuser/alcoholic/co	Resting-state EEG; 10 min EO, 10 min EC	LDA, SVM, MLP, Logistic Model Tree	45 subjects: 12 abusers, 18 alcoholics, 15 controls	Time-domain, frequency power	Absolute power Relative power	96%	97%	93%	0.97	67
An EEG-based functional *Mumtaz et al. (2021) [37]*	Binary classification of AUD	Resting-state EEG, 64 channels	SVM, Naive Bayes, Logistic Regression	60 subjects: 30 AUD, 30 controls (age-matched)	Spectral entropy, alpha/theta ratios	Synchronization Likelihood Interhemispheric Coherence Mutual Information (MI) Integration of SL, Coherence and MI	98%	100%	95%	0.97	68
Classification of EEG signals *Rodrigues et al. (2019) [38]*	Alcoholism vs. control	Online EEG; protocol not described	SVM, OPF, Naive Bayes, KNN, MLP	UCI KDD EEG dataset: 77 alcoholic, 45 control	Spectral band energy, statistical features	PCA	99.6%	99.6%	99.6%	–	64
Alcohol use disorder detection *Anuragi et al. (2020) [16]*	AUD classification	Online EEG; multichannel resting EEG	LS-SVM, SVM, Naive Bayes	UCI KDD EEG dataset: same as above	Statistical features, DWT features	Not specified	99.2%	100%	99.2%	0.992	82
Deep convolutional neural *Mukhtar et al. (2020) [39]*	Alcoholism detection (binary)	Online EEG; raw data input	CNN	UCI KDD EEG dataset: same as above	Raw EEG signals	Auto-learned via CNN	98.4%	–	–	0.984	45
Classification of alcoholic EEG *Farsi et al. (2020) [40]*	AUD classification: healthy vs. alcoholic	Online EEG; temporal deep learning	ANN, LSTM	UCI KDD EEG dataset: same as above	Time sequences, mean + STD + var	Not specified	93%	–	–	–	64

[1] ANN – Artificial Neural Network,
[2] AUD – Alcohol Use Disorder,
[3] CNN – Convolutional Neural Network,
[4] DWT – Discrete Wavelet Transform,
[5] EEG – Electroencephalography,
[6] EC – Eyes Closed,
[7] EO – Eyes Open,
[8] GEE – Generalized Estimating Equations,
[9] IGD – Internet Gaming Disorder,
[10] KNN – K-Nearest Neighbors,
[11] LDA – Linear Discriminant Analysis,
[12] LSTM – Long Short-Term Memory,
[13] MLP – Multi-Layer Perceptron,
[14] OPF – Optimum Path Forest,
[15] PCA – Principal Component Analysis,
[16] SVM – Support Vector Machine

Table 2.2: Summary of EEG-based studies on alcohol use disorder detection using machine learning, including experimental setup, signal features, and classification performance

3. Methodology

3.1 Overall methodology

In this section, two experimental platforms are used:

- National Survey on Drug Use and Health 2019 online database
- Brain-computer interface experiments using alcohol stimulus

to correlate the clusters of clinical assessments and alcoholism experiments using unsupervised machine learning. An outline of the overall methodology is given which describes the steps to be taken to achieve the desired results. For each experimental platform, a framework of the methods and materials used will be presented.

3.1.1 Outline

1. Gather and analyze data of clinical assessments based on DSM V
    - Consolidate online data of clinical assessments
    - Build three unsupervised machine learning models to cluster the data
    - Validate the accuracy of clustering
    - Confirm the findings with a mental health professional
2. Perform human experiments involving alcohol stimulus
    - Gather volunteers, preferably people who drink alcohol
    - Provide an increasing amount of alcohol at specific time intervals
    - Record EEG signals throughout the experiments
    - Cluster the results of recorded brain signals using unsupervised machine learning

3.1.2 Gather and analyze data of clinical assessments based on DSM V

In the methodology, the first step involves gathering online data of clinical assessment based on DSM-5 criteria for SUD. Data gathered from an online database provides a review of the

use of drugs and substances and as such, the volunteers' responses were consolidated. In the data, subjects were provided with questionnaires to better understand their frequency of use and dependence, therefore the study aims to discover the SUD status of individuals based on the responses. This will be achieved by using statistical methods and machine learning algorithms to find insights into the data and then cluster them into relevant categories. Three unsupervised machine learning models will be built to cluster the data and then verified.

3.1.3 Perform human experiments given alcohol stimulus

After going through a clinical assessment, seventeen male volunteers were gathered who consume alcohol. The subjects were ready for the experiment which involves recording brain signals when an alcohol stimulus is given that was provided at specific time intervals. The EEG data was used to build an unsupervised learning model to cluster the brain signals. The results were used to correlate the clinical assessments and EEG signals.

3.2 National Survey on Drug Use and Health (NSDUH) 2019 online database

The NSDUH survey contains clinical assessments of substance use disorders for different types of drugs and substances including illicit drugs, tobacco, and alcohol. It has publicly available datasets and is a source of statistical information [33] covering aspects of drug and substance use while in addition providing questionnaires from the DSM to assess mental health disorders associated with relevant addictive substances [44]. The questionnaire considered is based on the latest manual which is the DSM-5 and published in 2013. Data from N=56,137 subjects was collected for clinical assessment, where each individual had to go through a psychiatric evaluation to diagnose their level of SUD. The participants in the survey are adolescents aged 12 to 17 years old and adults aged 18 years or older in the United States [45]. The method of data collection was random, as the questionnaire was given to different families without prior knowledge of their exposure to illicit substances which means that their classification of having SUD or non-SUD is not known. This work considered variables from the sections self-administered substance use, risk availability for that particular substance, and substance abuse and dependence. The last measure was used to diagnose the patients to find their level of SUD and all the features have variable code

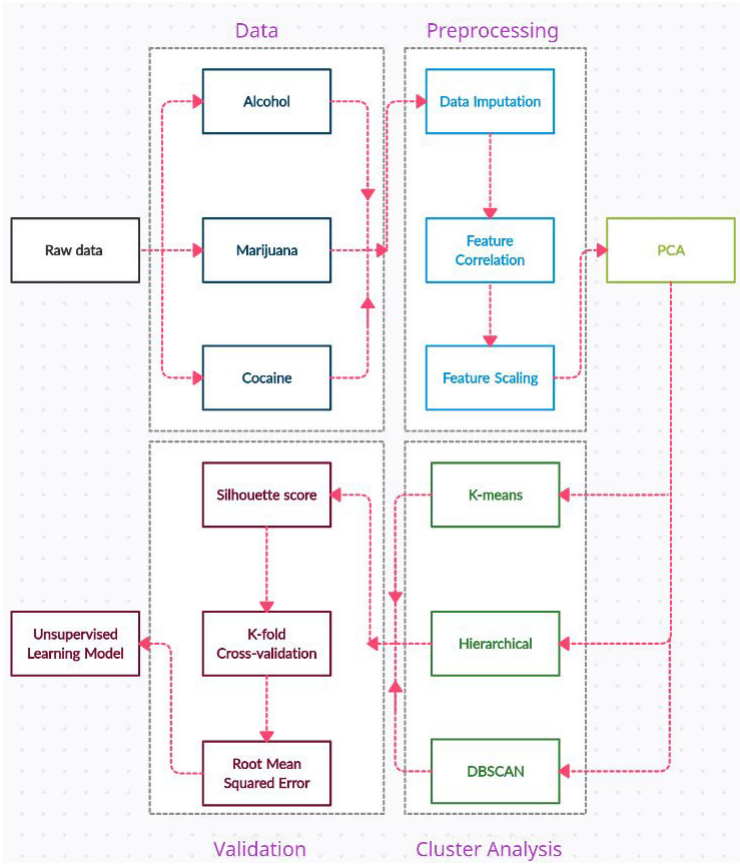


Figure 3.1: The framework showing the proposed model structure to analyze data from an online database. The arrows indicate the sequence of our approach where preprocessing, cluster analysis, and validation algorithms were applied consecutively for each dataset. There are three substances considered, namely, alcohol, marijuana, and cocaine. Three unsupervised machine learning algorithms are used, namely, k-means clustering, hierarchical clustering, and DBSCAN.

names. The study focuses on substance use disorder on three substances: alcohol, marijuana, and cocaine. The initial step identifies the input variables associated with the features that were used to represent patient responses during the assessment.

3.2.1 Data Preprocessing

Upon review of the large online dataset, it was noticed that the collection of data included ten types of drugs plus alcohol, and tobacco but the main focus of this study will be on alcohol, marijuana, and cocaine SUD. Manual extraction of relevant features was carried out in which answers to the questionnaire were based on DSM 5 criteria for substance use disorder and the same process was carried out for all three datasets. Two columns were dropped for participant identification which include their identification number and file date showing the data creation date. The pre-processing step is crucial in any machine learning study as it prepares data for modeling by removing errors to make data smooth, understandable, and

noise-free. It is also important in creating more accurate results by removing unwanted variabilities [46, 47].

3.2.1.1 Data Cleaning

This is a data treatment technique whose main purpose is to deal with missing data, to avoid both bias and overestimation of the prediction accuracy in psychiatric research [48]. Most of the features included missing data, indicating that some of the questions during the psychiatric assessment were not answered by participants while some dropped out of the survey [49]. It was observed that the values in the dataset were missing completely at random (MCAR) because the probability of missing data is not dependent on either observed or unobserved data [50]. A rounded mean data imputation technique was used which allows missing values to be replaced. This will address the error in the data because of missing information. This is the first necessary step in data preprocessing. This method uses the mean of each feature by calculating the average of all the values in that sample. Since the data type is MCAR, mean imputation is necessary to prevent bias and keep variability between the input labels constant. This will reduce the magnitude between the largest and smallest values of the sample.

3.2.1.2 Feature Correlation

Feature Correlation is performed on the raw data with complete information. This is the second necessary step to find correlations of features from the raw data. The method here uses a linear relationship between features which can help discard irrelevant and redundant data to improve computational time [51]. It was achieved by creating a Pearson correlation matrix which indicated that few of the features were highly correlated. This method measures the linear relationship between input variables and ranges between the values 1 and -1 using product-moment correlation coefficient [52].

In the legend of the graphs, the violet shade indicates a higher correlation of variables while the red shade shows the least correlation. The results of feature correlation are shown in Fig. 3.2, Fig. 3.3 and Fig. 3.4 sequentially.

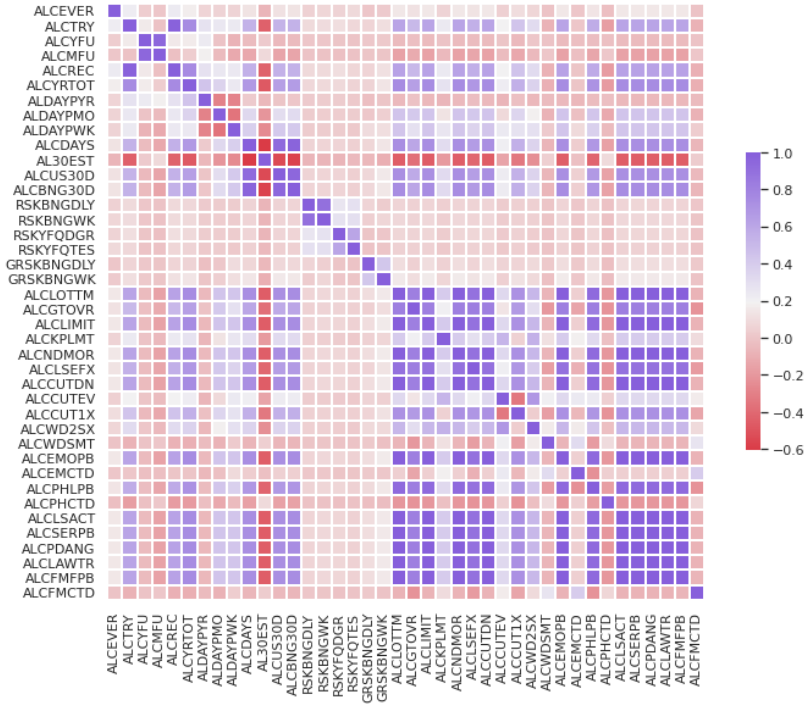


Figure 3.2: The figure shows the correlation matrix for the alcohol dataset. A more dominant and positive correlation can be observed from the 20th feature to the last feature which is depicted by the concentrated purple color.

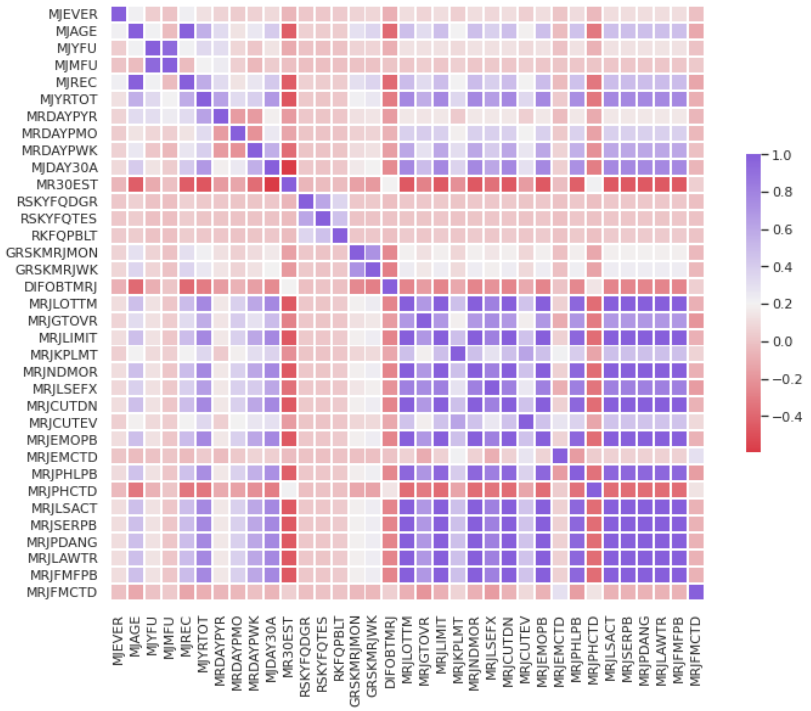


Figure 3.3: The graph shows the correlation matrix for the marijuana dataset. The graph is almost similar to the alcohol, except the features positively correlated can be observed from the 18th to the last feature.

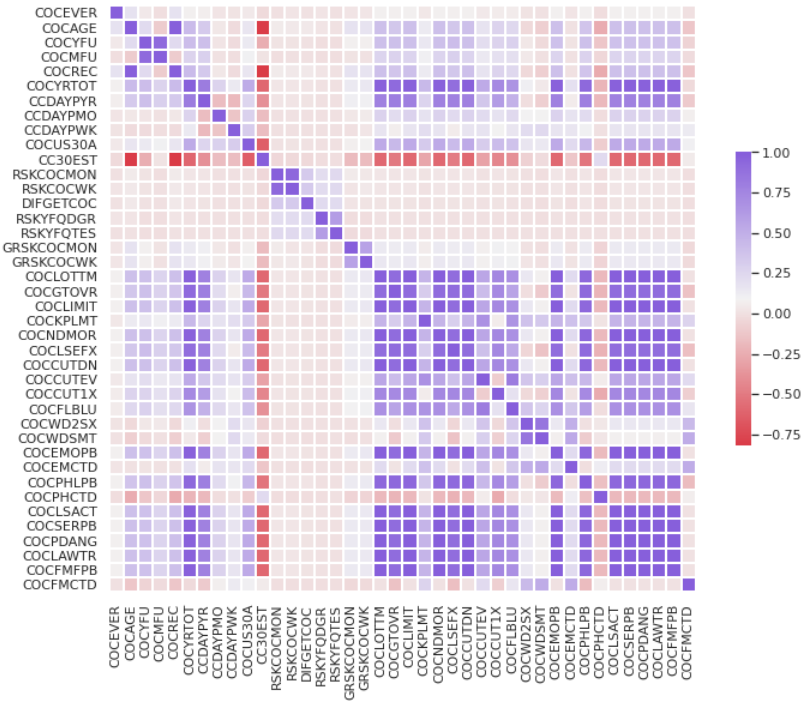


Figure 3.4: The figure shows the correlation matrix for the cocaine dataset. This graph also has similar characteristics to the alcohol and marijuana correlation matrices in the sense that the highly correlated features are those under the substance abuse and dependence section.

3.2.1.3 Feature Scaling

Feature scaling puts the data into a range of values that will equalize the strength of each data entry. This is the necessary third step that involves the computation of data values, with completed information and devoid of redundant or irrelevant data. This method will standardize the varying magnitude of the subject's responses during the survey [53], to scale down the minimum and maximum values to a common measure.

3.2.1.4 Dimensionality Reduction

In the survey, every patient response is weighted to derive numerical variables relating to the questionnaire. The range between different weights is really large, with a minimum value of zero and a maximum of nine hundred and ninety-nine depending on how the subject has answered the question. Dimensionality reduction deals with high-dimensional and multivariate data [54] to visualize a multi-dimensional space into a lower-dimensional space, e.g. 2D. This is the last necessary step in our data preprocessing to help us with data visualization by considering only the two most important components. There were 40 features for the alcohol and cocaine datasets and 35 features for the marijuana dataset.

3.2.1.5 Data-matching

Data matching is a process to establish the relationship among the input variables. It can help aid in identifying similarities or differences within the variables that could be overlooked due to assumptions about the data distribution [55]. In this work, the correlation between the features has to be minimized to reduce computational costs and to create the total extent of representation of the data space. The NSDUH database contains answers to questions from respondents that are converted to integers. Pearson correlation and PCA are implemented as data-matching techniques. Firstly, the Pearson correlation was utilized to establish the linear relationship between features as the data follows a normal distribution. This is because the majority of answered questions have a lower range and thus indicate that they are centered around the mean. The unanswered questions were assigned bigger weighted values to distinguish the relationship between different sets of responses. In this way, the Pearson correlation matrix was able to classify the features between positively correlated, no correlation, and negatively correlated. Secondly, PCA was used to identify the attributes according to their order of importance, that is, from the most important to the least important. This technique compresses the features and retains the most critical variables. This process utilizes eigenvalues of the identified attributes from the largest to the smallest value, in their order of importance. This data-matching technique allows us to select the two most important attributes to analyze the data in the most significant two-dimensional data space.

3.2.2 Cluster Analysis

A data mining technique that separates groups with dissimilar properties to analyze data by grouping similar points called clusters [56]. The clustering of unlabelled data was performed by the Python module called sci-kit-learn. The algorithms used examine and implement two types of variants, namely, the fit method and a function. The former learns the clusters on train data while the latter returns an array of labels that correspond to the different clusters when given train data [57]. Three clustering techniques are used to compare predictive accuracy, namely, k-means, hierarchical, and DBSCAN. After performing dimensionality reduction, the data was fed into these models to create 2-D visualization of the results. According to [58], the clustering techniques can be of three types, namely, partitioning, density-based, and hierarchical-based.

3.2.2.1 K-means Clustering

K-means is a partition clustering technique that is widely used due to its simplicity and remains one of the best distance-based algorithms [59, 60]. The sum of the squared Euclidean distance between the data points and the centroid has to be minimized to allocate them to a cluster [58]. From this principle, the internal coherence of the clusters can be measured. For a given dataset **X** containing $\mathbf{x}_i$ data points where i=1,...n, and n is the number of cluster labels. The sum of the squared error (SSE) can be minimized using the following equation:

$$SSE = \sum_{j=1}^{k} \sum_{i=1}^{n} ||\boldsymbol{x}_i - \boldsymbol{\mu}_j||^2 \tag{3.1}$$

where *k* is the number of the clusters, $\mathbf{x}_i$ is *i*-th data point, and $\boldsymbol{\mu}_j$ is the mean of the samples within the *j*-th cluster. We select *k* initial seeds and assign a set of points into k clusters by minimizing SSE. The elbow method is used to find the optimal number of clusters by choosing a threshold at the elbow or bend of the graph. It uses the idea of squaring the distance between sample points in each cluster and the centroid to generate a series of k values [61, 62].

3.2.2.2 Hierarchical Clustering

Agglomerative hierarchical clustering was used for this technique where each point is considered an individual cluster initially until similar clusters are merged to form one big cluster. Just like k-means, it needs a predefined number to obtain optimal clusters and that is achieved by the use of the technique called dendrogram that follows a bottom-up structure where each data point is initially an individual cluster. It keeps building a hierarchy by merging the closest clusters until there is only one cluster left. This algorithm calculates the Euclidean distance between the clusters and uses the linkage measure to establish which distance to use. The Ward linkage method narrows down the variance between the clusters. The Euclidean distance is expressed as

$$d\left(x, y\right) = \sqrt{\sum_{i=1}^{n} \left(x_i - y_i\right)^2} \tag{3.2}$$

where $\mathbf{x}_i$ and $\mathbf{y}_i$ represent the components of two points in an *n*-dimensional data space. We need to establish how far the points are from each other to be considered a cluster. The ward linkage can be expressed as

$$\Delta(A,B) = \frac{n_A n_B}{n_A + n_B} ||\overrightarrow{\boldsymbol{m}}_A - \overrightarrow{\boldsymbol{m}}_B||^2 \tag{3.3}$$

where Δ denotes the distance between the two clusters *A* and *B*, $\mathbf{n}_A$ and $\mathbf{n}_B$ are the number of data points in each cluster, $\mathbf{m}_A$ and $\mathbf{m}_B$ are the centers of clusters. A dendrogram method is used to obtain the optimal number of clusters for hierarchical clustering. All the clusters are co-joined until a complete structure is observed at the highest level.

3.2.2.3 DBSCAN

DBSCAN algorithm relies on the density of data-point distribution [63]. It uses two parameters epsilon (eps) and minimum points (minPts) to identify the density of data and then classify data points as either core point, border point, or an outlier. The eps specifies the closeness of points to be considered a cluster and thus any point with a distance lower or equal to the value of eps will be considered a neighbor. The minPts are the required minimum points to form a cluster. Any data point which is not a core point or border point will be considered noise. The core point has to contain at least minPts and the border point must be within the neighborhood of the core point. For a dataset *D*, a data point *p* can be located that belongs to its data space. To obtain the distance between *p* and its neighbors, we define a density distance *d* that computes the ratio between the maximum density distance d_max and the minimum density distance d_min of the points from *p* [64]. The equations for these parameters can be expressed as

$$d_max(p) = max\{d(p,q) \mid q \in D \wedge d(p,q) \leqslant eps\}$$
$$d_min(p) = min\{d(p,q) \mid q \in D \wedge d(p,q) \leqslant eps\}$$

where *q* is a data point in the neighborhood of *p* that belongs to *D*. To obtain the values of *d*_max and *d*_min, *q* should satisfy the conditions that it is an element of *D* and that the distance between *p* and *q* is less or equal to the eps.

Admittedly, this study used three basic algorithms that have been used for many years now. This is because, they prove to be useful in clustering clinical data as shown in the following recent references: [65, 66], and [67].

3.2.3 Validation

To confirm the cluster results, this work incorporated metrics in the study that checked for the prediction accuracy and error deviation of the three algorithms. It used silhouette score for internal validation and k-fold cross-validation, and root mean squared error for external validation.

3.2.3.1 Internal validation

The objective is to assess the quality of the clustering structure between the data itself. Internal validation can also help choose the best clustering algorithm as it relies on the information presented in the data [23, 68].

Silhouette score is used to evaluate clustering quality and how clusters compare to each other within the algorithm. The values range between negative and positive one. Results are outlined as percentage values in which scores close to one are distinguished but model performance decreases towards a negative one.

3.2.3.2 External Validation

The quality of the clustering is assessed by comparing observed cluster labels with predicted labels [23]. The main objective is to find out if a model can correctly classify the test variables with similar characteristics. THis works evaluates the prediction accuracy of the algorithms with a new set of data. A comparison is made to the observed variables to determine the derived prediction rule [69]. The 10-fold cross-validation is used to analyze the prediction accuracy of each algorithm and test the model performance for new data.

The k-fold cross-validation approach evaluates the performance of a classification model. However, it was deployed on the clustering algorithms by using class labels as target variables. It works on the principle of separating the dataset into k-folds in which $k - 1$ folds are used as a training set and the remaining is used as a test set. Depending on the value of k, the process will iterate through that value until all the folds have been used for testing [70, 71]. The data is partitioned into ten folds which were randomly selected and part of them were used as the training set while the remaining was an evaluation set. The accuracy of each fold was calculated and the average for the 10 folds was evaluated. The results are presented in Table 4.1. The RMSE technique is the square root of the mean squared error which measures the error generated between observed and predicted values. This metric was used to calculate the difference between observed and predicted values for cluster labels. After performing standardization, the majority of values were centered around the mean to give a Gaussian distribution. This metric is appropriate for error prediction for such cases [72, 73].

A study closest to our work is [23] in which some computational techniques are the same as theirs. They are as follows: PCA for dimensionality reduction, k-means for clustering, and silhouette score, k-fold cross-validation, and RMSE for validation. The silhouette scores for [23] are 76% for emotions, 82% for chronic disease, and 69% for yeast. The silhouette scores are as follows: 72.5% for alcohol, 78.7% for marijuana, and 63.9% for cocaine using k-means clustering. The discrepancies in the results can be attributed to differences in data preprocessing where [23] used Calinski–Harabasz index, while our work used mean imputation, feature correlation, and standardization. However, the main difference is the data source, where [23] used emotions, chronic diseases, and yeast datasets, while this work used the NSDUH dataset. In a sense, the validation scores are within the acceptable range in both approaches despite the difference in the source datasets.

3.3 Brain-computer interface experiments using alcohol stimulus

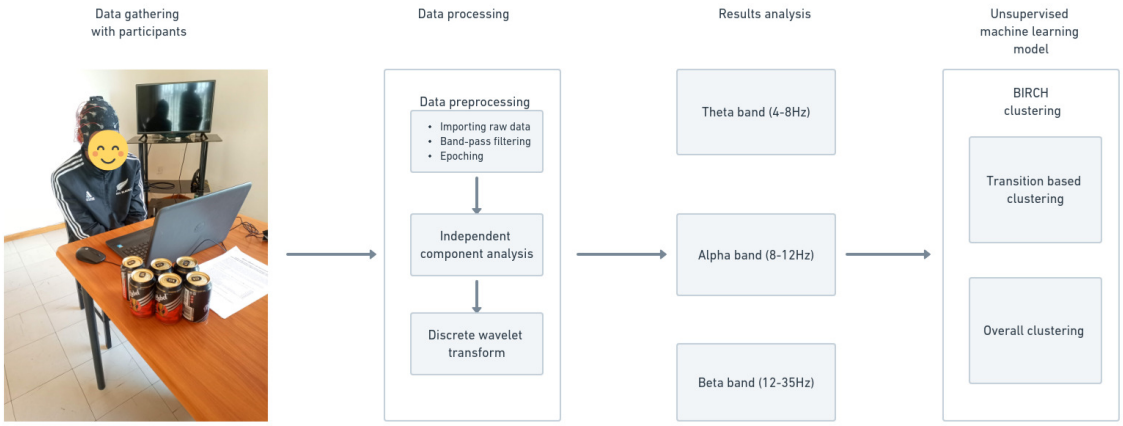


Figure 3.5: The block diagram showing the different steps taken to analyze the raw EEG data. The arrows indicate the sequence of our proposed framework, starting with the recording of brain signals, followed by processing, and results analysis. We use the BIRCH algorithm to make a classification for transition-based and overall clustering results.

3.3.1 Data gathering with participants

3.3.1.1 Demography of participants

In this work, data was collected from a total of seventeen alcohol users recruited through oral communication and social media. The study is positioned as a pilot study to evaluate the feasibility of applying unsupervised machine learning to EEG experiments data involving alcohol stimulus. While

no formal sample size calculation was conducted, similar exploratory EEG studies have used sample sizes in the range of 15–30 participants [17, 74], supporting the suitability of this sample for preliminary analysis. The study population consisted of only male subjects with an age range of 22 to 44 years and a mean age of 27.88±7.08 years. Only individuals with a history of alcohol use could be included in the experiments and free alcohol was provided to any person willing to participate in the study. Participation in this study was voluntary and each subject was briefed about the experimental procedure including eligibility criteria, purpose of the study, privacy, anonymity, and confidentiality. Information on the identity of the participants has been kept confidential and will not be made public. All volunteers had to sign the consent form before taking part in the alcoholism experiments.

3.3.1.2 Eligibility criteria

According to the eligibility criteria as stated in the consent form, any male aged 18 to 60 years with no history of mental illnesses was allowed to participate in the experiments. At the time of the examination, the volunteers were advised to have very short, dry, and un-oiled hair to increase the contact between the electrodes and the scalp. To be eligible, the participants had to answer **Yes** to question 1 and **No** to questions 2 to 9.

1. Have you ever consumed 2 liters of 6% alcohol or more in two hours before?
2. Do you have any alcohol allergies?
3. Have you ever had a bad reaction to alcohol?
4. Do you suffer from any liver disease?
5. Do you suffer from diabetes?
6. Do you have any history of diabetes?
7. Do you suffer from kidney disease?
8. Do you have any history of kidney disease?
9. Have you ever been diagnosed with Alcohol dehydrogenase deficiency?

3.3.1.3 Self assessment

Each volunteer was given an AUDIT to test their frequency of use of alcohol and their susceptibility to AUD. The AUDIT is a standard screening questionnaire that identifies individuals at risk of alcoholism. This self-report tool comprises 10 questions about possible drinking patterns and the related

consequences. The total possible scores for this test range from 0-40 with higher scores showing an increased probability of severe drinking patterns. Possible responses to the questions are within the range of 0-4 except for questions 9 and 10 [75, 76]. According to the guidelines by the World Health Organization (WHO), a score of 1-7 indicates low-risk alcohol use while a score from 8-14 suggests harmful use of alcohol. Similarly, scores ranging from 15-40 indicate the likelihood of AUD, that is, moderate to severe alcoholism.

3.3.1.4 Experimental Device

The device used in the experiments is a 32-channel non-invasive EEG wireless headset with a sampling rate of 128Hz. A gel sensor-based Emotiv Epoc Flex cap shown in Fig. 3.6 was used to record the brain signals of participants under the influence of alcohol. It is used for advanced BCI applications because the electrodes allow for easier and more accurate brain signal acquisition. This device provides flexibility and an adjustable design to accommodate different head sizes. In the alcoholism experiments, the sensors were placed following the 10-20 standard electrode placement. The EEG signals are recorded from 32 channels that include: Cz, Fz, Fp1, F7, F3, FC1, C3, FC5, FT9, T7, TP9, CP5, CP1, P3, P7, O1, Pz, Oz, O2, P8, P4, CP2, CP6, TP10, FC6, C4, FC2, F4, F8, and Fp2 including two reference sensors. In Fig. 3.7, each electrode is labeled using a channel name and they are marked in green color while the reference electrodes are marked in black color.



Figure 3.6: The Emotiv EPOC Flex Gel Sensor 32-channel EEG headset used to harness the brain activity of participants during the BCI experiments with alcohol stimulus

3.3.1.5 Experiment procedure

Before commencing the alcoholism experiments, a registered nurse recorded and monitored the vital signs of the volunteers to declare their fitness to participate. All eligible subjects were given a six-pack of alcohol each that contained either 5.5% or 6% of alcohol content depending on preference.

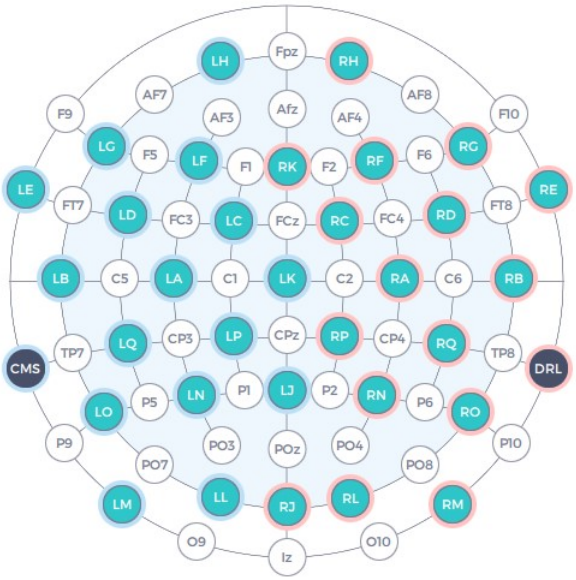


Figure 3.7: A figure showing the montage standard electrode placement for a 32-channel EEG headset that follows the 10-20 system

The duration of the experiment was two hours for each volunteer and the brain signals were recorded throughout the entire period using a non-invasive EEG head module. The rate of consumption was based on the willingness of the volunteer to take the alcohol but the timing and estimated amount were confined within the two hours. To assist in restraining any possible violent behavior arising from the effects of alcohol on the volunteers, two male security officers were present during the experiments. No adverse side effects were observed in all subjects.

3.3.2 Data preprocessing

This study's raw data harnessed during the alcoholism experiments was first imported in comma-separated variables (CSV) format. Filtering was applied using notch and band-pass filters to remove artifacts in the data. The data was then segmented into 20-minute intervals for epoching and creating fixed-length events. ICA was used to separate independent components from the raw data using MNE [77]. Finally, DWT was applied to decompose the ICA-processed data into time-frequency domain features.

3.3.2.1 Importing raw data

The initial step in the EEG analysis was importing the raw data recorded during the experiments. There are different types of storing EEG data which are Functional Imaging File (FIF), European Data Format (EDF), and (CSV). In this study, the CSV format was used because it allows for easier manipulation and analysis of brain signals. A diagram of the raw EEG signals in continuous time is

shown in Fig. 3.8

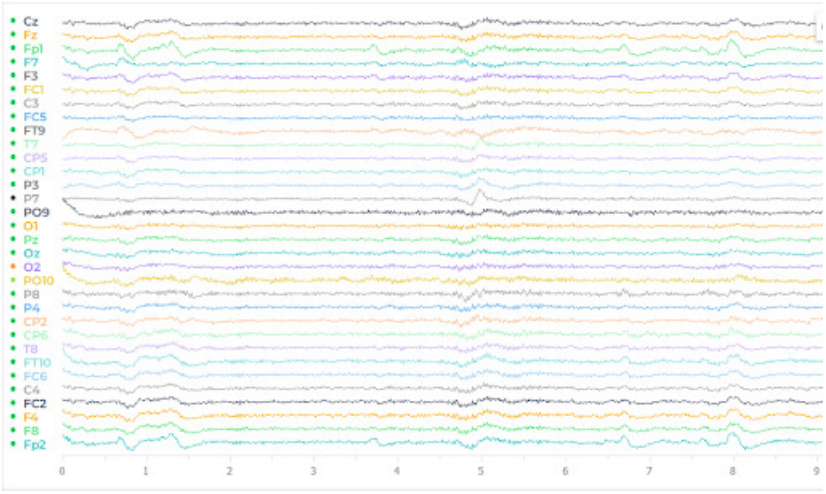


Figure 3.8: The raw EEG signals as recorded from the brain of a subject during the alcoholism experiments. It shows the 32 channels and their activity in continuous time

3.3.2.2 Filtering

To reduce physiological and non-physiological artifacts and improve the quality of the data, notch and band-pass filters were implemented. Brain signals are usually affected by power line noise at 50 or 60Hz depending on the region [78]. To remove this electrical interference, a notch filter was used in the data with a cutoff frequency of 60 Hz. The second artifact removal step is a band-pass filter that was implemented within the range of 0.5 to 55Hz. This type of filter allows for the isolation of the delta (0.5-4 Hz), theta (4-8 Hz), alpha (8-12 Hz), beta (12-35 Hz), and gamma (>35 Hz) brain rhythms to enhance the signal-to-noise-ratio of the data [79].

3.3.2.3 Epoching

This step involves creating time-locked data segments called epochs. In this study, the transitions that occur in the brain when a person drinks alcohol are observed. As such, the data was segmented into 20-minute epochs where each segment represents one bottle of alcohol intake. For every subject, the total number of epochs was six and each represented a bottle of alcohol intake.

3.3.2.4 ICA

This is one of the most prominent blind source separation techniques in which independent components are estimated from a set of EEG recordings. The mixed signals such as eye blinks and muscle movements are removed to ensure that analysis of the EEG data is independent of the external in-

terference [78, 80]. ICA provides a more artifact-free characterization of the brain by eliminating undesired segments from the original signal [81].

3.3.2.5 Discrete wavelet transform

Due to the non-linearity and non-stationary properties of the EEG signals, the extraction of the relevant features needed for the analysis may be complex [82]. A more robust method is needed to achieve accuracy in extracting informative characteristics of the data. Therefore, DWT has been proposed in this work to decompose the signals into a highly efficient representation called wavelets. This technique helps in capturing useful time-frequency domain features of the data at different resolutions [83, 84].

3.3.3 Dimensionality reduction

Due to a lot of redundant information in high dimensional data, this method can be utilized to eliminate those by converting them from high to low dimensional feature space [85]. In this study, the PCA technique was implemented to perform dimensionality reduction.

3.3.3.1 Principal Component Analysis

PCA is a widely used dimensionality reduction technique that aims to find the optimal variance from a set of features. These are transformed into smaller features that contain the most information about the data. It also reduces the complexity of a model while improving its efficiency. In the case of duplicate features, PCA can overcome this as it creates a new set of non-linearly related features [86]. It was applied to the DWT-processed data by defining two principal components to plot our clustering results in 2-dimensional feature space.

3.3.4 Clustering

This is an unsupervised learning technique in which a set of unlabelled data is analyzed based on the data correlations. Many different algorithms can be applied which are dependent on the type and size of the dataset [57]. In this work, the BIRCH method was used to cluster the PCA-processed data as it remained the best-performing algorithm compared to others.

3.3.4.1 BIRCH

This algorithm is a hierarchical clustering method that creates a tree-like structure using clustering features (CF) and cluster feature tree (CF Tree). It is memory-efficient and can handle large datasets

with a faster execution time [57, 43]. Compared to other clustering metrics, it had a faster execution time.

4. Results and Discussions

4.1 Clinical assessments for SUD

In this section, the graphical visualizations of the results from the three clustering algorithms are presented. This work starts with k-means, followed by hierarchical clustering, and DBSCAN. The first and second algorithms required an additional technique to define the number of clusters before plotting the data points. The clustering results showed similar visualizations across the three algorithms in classifying SUD for the three substances, and are validated using silhouette score, k-fold cross-validation, and RMSE.

4.1.1 Alcohol dataset

For the alcohol dataset, the model determined that the two optimal number of clusters is two, as shown in Fig. 4.1. This is because the elbow method showed a big change in the slope of the graph. A big downward slope is shown in one to two clusters, while there is a slight downward slope from two to nine clusters making two clusters the "elbow" of the slopes.

4.1.1.1 K-means clustering

This study proceeds to show the k-means clustering using two clusters as shown in Fig. 4.2. The structure shows very tightly spaced data points within each cluster while some few are scattered around which also indicates a larger Euclidean distance from the mean. The red cluster is centered around -4 in the pca1 axis and in the range of -2 to 4 in the pca2 axis. The boundary between the red and the green cluster is roughly around 0 of the pca1 axis. The region of the red cluster is bounded by the range -6 to 0 in the pca1 axis and -4 to 10 in the pca2 axis. The green cluster is less dense than the red cluster and each region is bounded by 0 to 8 in the pca1 axis and -4 to 4 in the pca2 axis. It is noted that the green cluster has a longer range along the pca1 axis compared to the red cluster, while the red cluster has a longer range along the pca2 axis compared to the green cluster.

The red cluster being denser within its region clearly states that the red cluster is most likely classified correctly. The more points present in the same region support the consistency of the computed results, thus confirming the commonality of answers by several respondents. The green cluster is less dense compared to the red cluster, but its denser region can be virtually recognized from the

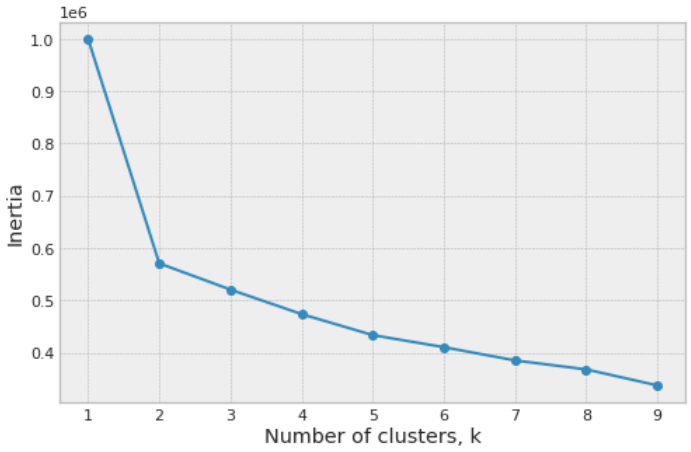


Figure 4.1: The figure represents the elbow method under the alcohol dataset to determine the optimal number of clusters. The graph shows a bend at k=2, indicating two clusters in the dataset. A smooth line immediately after the inflection point can be observed, which is an indication that it is where our model fits best.



Figure 4.2: The k-means clustering for the alcohol dataset showing two distinct groups as per the elbow method. Some data points are closer to the boundary across the groups but have been assigned to their respective clusters. It indicates that the model chose the lesser Euclidean distance for each point. Overall, the red cluster has more points to show the majority of subjects belong to one group.

graph. This is because the classification of SUD can be categorized into two major classifications, with one classification consisting of the majority of the respondents. It is noted that the region of the green cluster is around two-thirds of the size of the red cluster. This is because the questions given to the respondents are uniform, but the answers depend on the respondents' experience. Thus, the cluster regions are defined by the features that are based on the answers to the survey. However, the green region is smaller because it consists of fewer respondents whose answers can be less diverse compared to when the respondents are much more.

4.1.1.2 Hierarchical clustering

For hierarchical clustering under the alcohol dataset, it was determined that there are two optimal clusters as shown in Fig. 4.3. In the model, the x-axis represents the data points distribution while the y-axis represents the cluster distance. The dendrogram shows a hierarchy with two main blocks of classification joined by a line of the height of around 900. The huge value of the connecting line means that the distance between the two classifications is big. This shows that indeed there are two distinct major classifications for the answers from the respondents. The orange block represents the first cluster and the green block represents the second cluster.

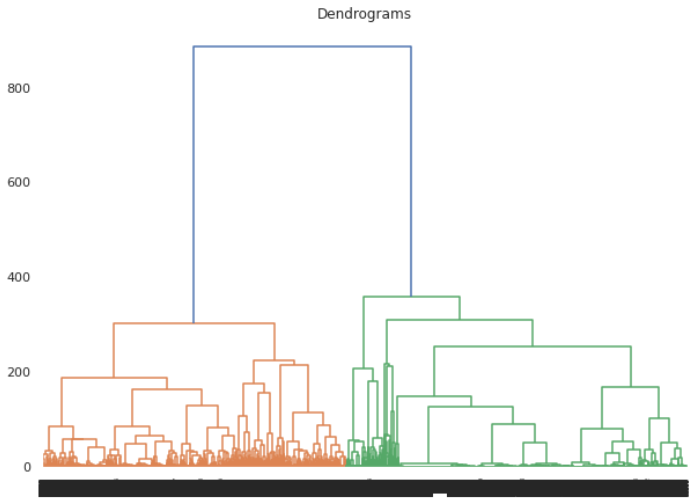


Figure 4.3: The dendrogram for the optimal number of clusters under hierarchical clustering is shown in the figure. The structure of the hierarchy shows an approximately uniform distribution with two clusters, however, we notice that the first group has more congregated data than the second one.



Figure 4.4: The hierarchical clustering for the alcohol dataset is shown in the figure. Results of the clustering show that the first group has more data points than the second and the model confirms the two clusters analogy by the dendrogram.

In the graph, the data points in the orange cluster are closer to each other and have shorter cluster

distances within the range of 0 to 300. This implies that the data points are dense and well-distributed throughout the entire range. It can also be said that the orange cluster consists of a much more number of respondents compared to the green cluster. For the green cluster, the data points near the boundary between the clusters are denser but dispersed further along the x-axis. The cluster distances are within the range of 0 to approximately 375. This means that the distances between the data points in the green cluster are bigger compared to distances between points in the orange cluster and thus confirms that the orange cluster is much denser than the green cluster.

These are results from hierarchical clustering that was used to classify two clusters as shown in Fig. 4.4. The structure is almost similar to that of k-means except that on the pca1 axis, the orange cluster is bounded in the region -6 to approximately 2, but in k-means it was -6 to 0. The boundary between the orange and the blue cluster is somewhere between 0 and 2 where a clear separation between clusters can be recognized. On the y-axis, the orange cluster is bounded in the region -4 to 10 while the blue cluster ranges from -3 to 4. These boundaries are identical to the pca2 axis boundary for k-means. The orange cluster has more data point distribution than the blue cluster and confirms the results obtained by the dendrogram, as well as the k-means. We can say that the results of hierarchical clustering are very similar to the k-means clustering. The only visible difference is the boundary along the pca1 axis between 0 and 2.

4.1.1.3 DBSCAN

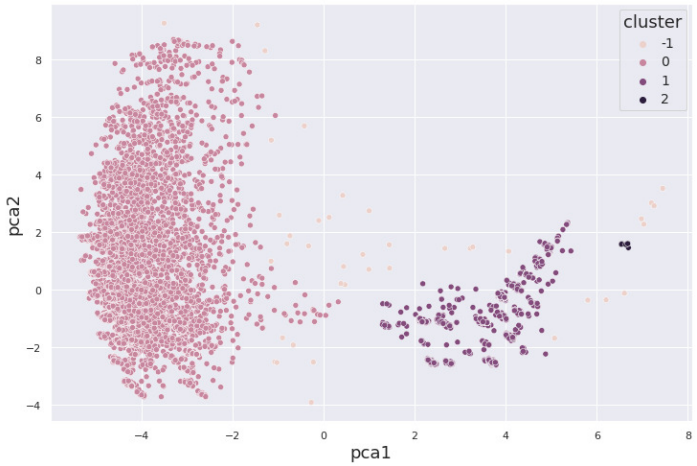


Figure 4.5: DBSCAN results for the alcohol dataset showing three distinct clusters with noise. The structure of the data points distribution has not changed for all three algorithms. Our analysis indicates that we have three levels of AUD for this algorithm.

For DBSCAN, the graph visualization, in general, showed similar results to k-means and hierarchical clustering. The centroid of the clusters and their boundaries are roughly similar in all three methods as shown in Fig. 4.5. However, in the results shown, there are four clusters with two ma-

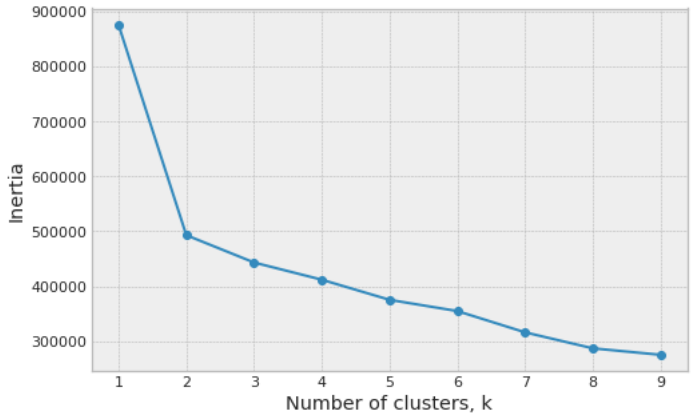


Figure 4.6: The elbow method under the marijuana dataset with a smooth line after the elbow of the graph to define two clusters for visualization.



Figure 4.7: The k-means cluster results for the marijuana dataset are shown. The few points between the clusters have greater Euclidean distance from the centroid but still form part of the respective clusters. More respondents are belonging to the red cluster than the blue cluster.

jor clusters that are similar in form to k-means and hierarchical clustering. The two minor clusters labeled -1 and 2 are lying away from the centroids of the two other major clusters labeled 0 and 1. However, one can say that if we choose to have only two clusters in the DBSCAN, then the results would be very much similar to k-means and hierarchical clustering.

4.1.2 Marijuana dataset

For the marijuana dataset, it was determined that the optimal number of clusters is two because of the change in the slope of the graph in Fig. 4.6 at k=2. A big and smooth downward slope can be observed from k=1 to k=2 and another downward slope from k=2 to k=9.

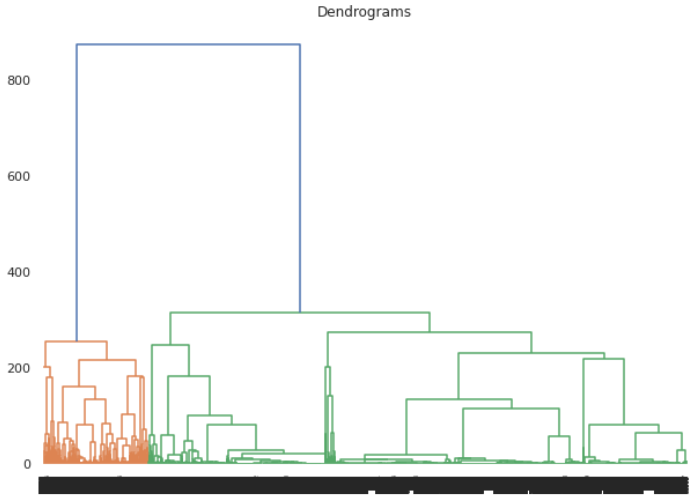


Figure 4.8: The dendrogram shows the density of points between each cluster. Based on the graph, we have two unique visualizations and the green cluster is more widely spread than the orange cluster.

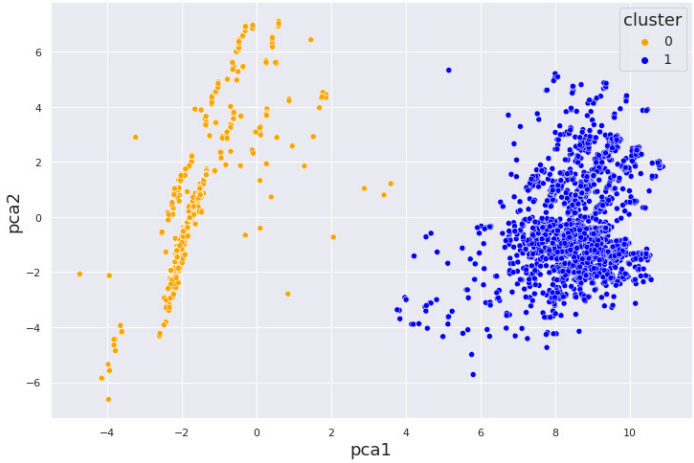


Figure 4.9: Hierarchical clustering results for the marijuana dataset are shown. Unlike for k-means, the two clusters are not bounded around 0 and the data points within their boundary have been assigned to the orange cluster in this graph.

4.1.2.1 K-means clustering

The sum of the squared distances from the centroid for the elbow method is given as inertia and plotted on the y-axis. It is noticed that there are much higher values of inertia on the graph compared to the other elbow technique graphs. This indicates a larger distance of the data points within a cluster and a much slower convergence in terms of allocating data points to their clusters.

The k-means clustering results for the marijuana dataset are shown in Fig. 4.7. The red cluster shows more data point distribution with a majority of dense data points and is bounded by the region between 3 and 11 in the pca1 axis, and from -6 to 6 in the pca2 axis. The green cluster has some dispersed data points bounded in the region from -6 to 3 in the pca1 axis and from -7 to 8 in the pca2 axis. This algorithm finds it difficult to deal with outliers, hence why they end up forming part

of the clusters. As observed on the graph, a couple of data points are far from the centroid yet still classified under respective groups.

4.1.2.2 Hierarchical clustering

The graph in Fig. 4.8 shows the dendrogram for the marijuana dataset. There are large intra-cluster distances and fewer data point densities on the orange cluster compared to the green cluster. The first data points are closely correlated but the distances increase as we move up the hierarchy and it is bounded in the region from 0 to approximately 250. The green cluster shows more data point distribution and density with a wider range. Hierarchical clustering results for the marijuana dataset are shown in Fig. 4.9. In the figure, the clusters are clearly distinguished. The orange cluster has dispersed data points with larger intra-cluster distances as only a few data points have minimal separation. It is bounded within the -7 to 8 region in the pca2 axis and from -5 to 4 in the pca1 axis. The cluster is less dense than the blue cluster which has more data point distribution, and lesser dispersion. The blue cluster ranges from 4 to 11 in the pca1 axis and from -6 to 6 in the pca2 axis. It can be said that this data is classified between two unique classes with the majority of respondents belonging to the blue cluster. The results of hierarchical clustering show very similar results to that of k-means.

4.1.2.3 DBSCAN

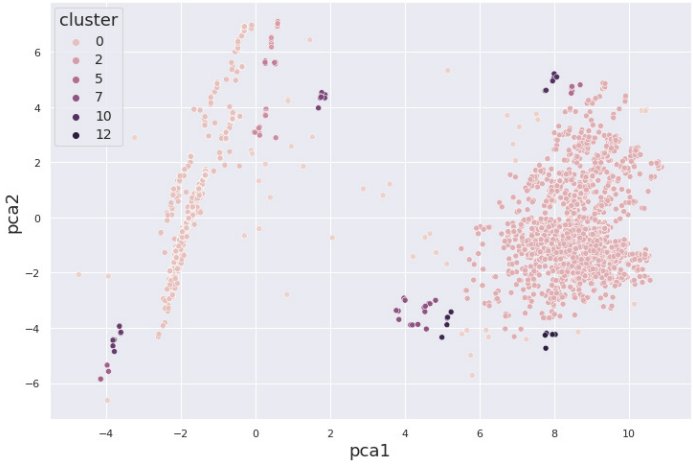


Figure 4.10: DBSCAN for the marijuana dataset is shown in the figure. There is no noise cluster meaning that we have no outliers in the dataset. There are many different groups, and the majority of data points belong to cluster 2.

The DBSCAN results for the marijuana dataset are shown in Fig. 4.10 with two major clusters and four other minor clusters. In general, cluster classification is similar to k-means and hierarchical

clustering. The only difference is that the four minor clusters with centroids lying far from the centroids of the two major clusters can be absorbed by them resulting only in two major clusters.

4.1.3 Cocaine dataset

4.1.3.1 K-means clustering

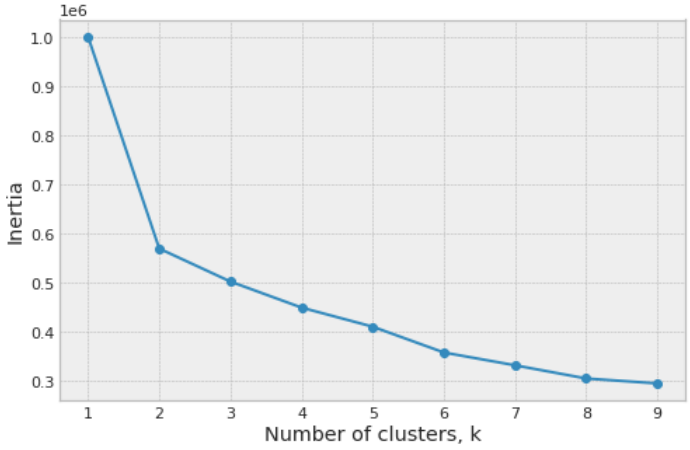


Figure 4.11: The elbow method for the optimal number of clusters which is almost similar to the previous datasets. We obtain a very smooth line immediately after k=2 to indicate two clusters for visualization.

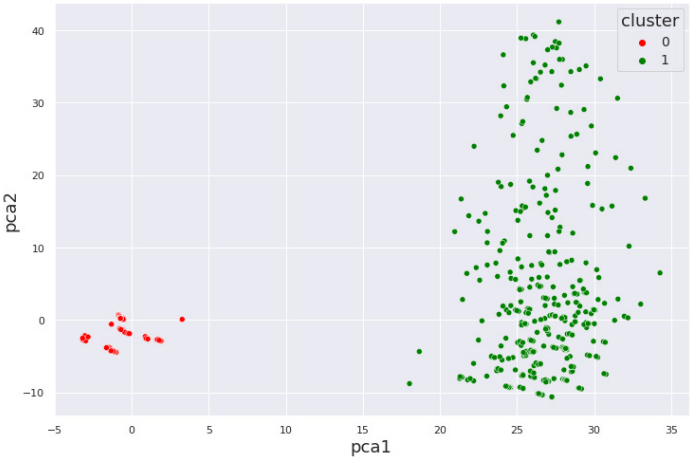


Figure 4.12: The k-means clustering results for the cocaine dataset are shown with well-spaced clusters. The intra-cluster distances however are more spaced in the green cluster.

The k-means clustering results for the cocaine dataset are shown in Fig. 4.12. There is a clear distinction between the two clusters and a much greater Euclidean distance. The red cluster indicates a close relationship between its members as most of the data points are located around the centroid at a much lesser distance. Bounded in the smallest region, the red cluster has the least data point distribution between -5 to 5 in the pca1 axis and -10 to 0 in the pca2 axis. For the green cluster, the majority of the data points are far from each by some degree but still share similar characteristics.

It is bounded within the range of 18 to 35 in the pca1 and -10 to 40 in the pca2 axis. The major difference between these results and the previous results for alcohol, and marijuana is that the clusters occupy a large scope/range of values in the pca1 and pca2 axis. Their values range from -6 to 10 and for cocaine, they range from -10 to 40 in the pca1 axis and -5 to 35 in the pca2 axis.

4.1.3.2 Hierarchical clustering

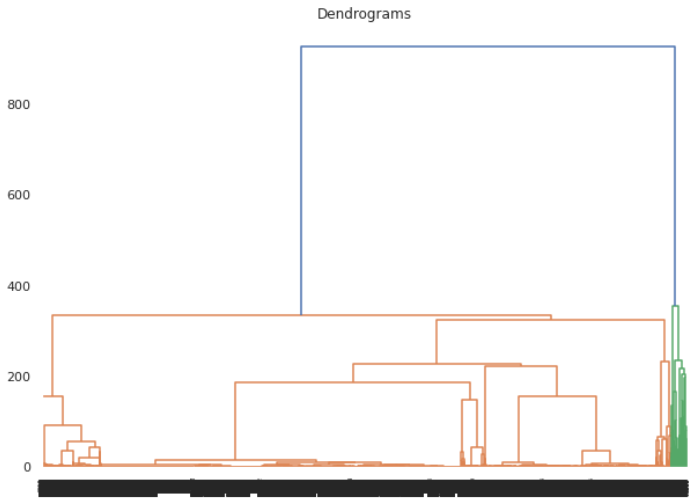


Figure 4.13: The dendrogram shows a major difference between the two clusters as the majority of data points belong to the first cluster. The figure also indicates that points in the second cluster should be closely distributed.

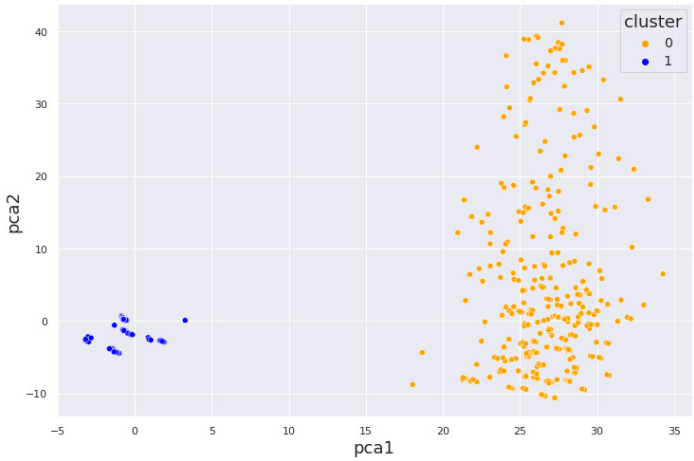


Figure 4.14: Hierarchical clustering for the cocaine dataset with a high density of points in the first cluster. Although we expected a close relationship between the data points, in the figure we can observe a separation between some points with one not forming part of the subgroups. This structure is similar to that of k-means in Fig. 4.12.

The graph for the dendrogram under the cocaine dataset is shown in Fig. 4.13 with two optimal clusters. A very close correlation between data points in the green cluster is shown while those in the orange cluster occupy a wider range of values.

The graph for the elbow method under the cocaine dataset is shown in Fig. 4.11. It was determined that there are two optimal clusters as indicated by a bend at k=2.

Compared to the elbow method under the marijuana dataset, the convergence for the elbow method under the cocaine dataset was much faster as indicated by the small inertia values. The values are also similar to the elbow method under the alcohol dataset.

The hierarchical clustering results for the cocaine dataset are shown in Fig. 4.14. The cluster classification for this algorithm is similar to k-means clustering in terms of data point density per cluster.

4.1.3.3 DBSCAN

The DBSCAN results for the cocaine dataset are shown in Fig. 4.15 with two major and four minor classifications. The clustering structure is similar to k-means and hierarchical clustering, but the difference is that it has overlapping data points and groups. Overall, six clusters are plotted on the graph and they are labeled from 0 to 25 with a step size of 5. The minor clusters formed within these main clusters and their centroids are closer to the centroids of the main clusters.

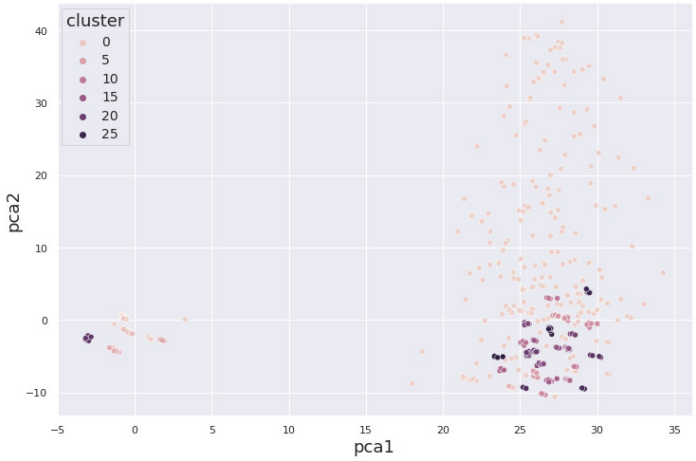


Figure 4.15: DBSCAN for the cocaine dataset with many clusters assigned very close to each other.

4.1.4 Validation Results

Table 4.1 presents the validation scores obtained for each algorithm. The performance of the three techniques is tested for each dataset and the results are outlined as percentages. These metrics were used to assess the cluster relationship, prediction accuracy, and error deviation.

In the following paragraphs, the study discusses the internal and external validation results obtained using the classical methods.

Algorithm	Dataset		
	Alcohol (%)	**Marijuana (%)**	**Cocaine (%)**
K-Means			
Sil_score	72.5	78.7	63.9
K-fold	99.96	99.92	99.84
RMSE	4.0	0.0	6.0
Hierarchical			
Sil_score	69.2	83.4	47.5
K-fold	99.97	99.96	99.63
RMSE	6.0	0.0	35.89
DBSCAN			
Sil_score	97.7	97.6	93.1
K-fold	99.96	100.0	98.8
RMSE	0.0	0.0	15.16

Table 4.1: Internal and External validation scores: Classical algorithms

Algorithm	Dataset		
	Alcohol (%)	**Marijuana (%)**	**Cocaine (%)**
K-Means++			
Sil_score	78.7	83.4	95.56
K-fold	99.99	99.99	100
RMSE	0.0	0.0	0.0
BIRCH			
Sil_score	68.7	61.2	96.6
K-fold	99.6	99.87	98.8
RMSE	8.9	8.0	15.12
HDBSCAN			
Sil_score	38.2	23.7	37.1
K-fold	93.7	92.3	86.4
RMSE	7.5	19.5	57.4

Table 4.2: Internal and External validation: State-of-the-art algorithms

The silhouette score obtained for k-means under alcohol dataset was 72.5% which means that the clusters are separated and data points are assigned satisfactorily. Based on the prediction accuracy and error, the performance of the model was very good as it managed to predict 99.96% of the possible assignment of clusters for our test dataset with an error of 4%. There is a percentage reduction for the silhouette score of 69.2% under hierarchical clustering compared to k-means while the error deviation increased by 2% as referenced in Table 4.1. The prediction accuracy obtained was 99.97% which remained the highest recorded for the three algorithms. The validation results for DBSCAN under the alcohol dataset showed very impressive values for the three metrics, as it obtained 99.96% and 0% prediction accuracy and error with a silhouette score of 97.7%. Generally, the performance of the clustering algorithms was much better under marijuana, as it recorded silhouette scores of 78.7% for k-means, 83.4% for hierarchical clustering, and 97.6% for DBSCAN. The prediction accuracy was 99.92% for k-means, 99.96% for hierarchical clustering, and 100% for DBSCAN while the prediction error remained 0% for all three algorithms. Cocaine recorded the lowest silhouette scores for k-means and hierarchical clustering which are 63.9% and 47.5% respectively. It also recorded the highest prediction errors of 6% and 35.89% for the two algorithms. For DBSCAN, it obtained a silhouette score of 93.1%, prediction accuracy of 98.8%, and prediction error of 15.16%. These results indicated the worst performance overall under cocaine for intra-cluster correlation and error deviation. The only approximately consistent scores are the prediction estimates for the model in which all of them were at least 99.9% as presented in Table 4.1. This is discussed in the following paragraph, where a comparison of the internal and external validation results between the classical and state-of-the-art methods is made. As can be seen from the tables, the state-of-the-art algorithm results shown in Table 4.2 are not far from the results of the classical algorithms. It is notable that for the most part, the classical methods perform better than the state-of-the-art. The interpretation is that state-of-the-art algorithms do not necessarily result in better performance and that the results can also be dependent on the type of data that is considered. In this case, the data gathered from the 25,000 respondents in the survey related to SUD proved to be processed better by the classical methods. Lastly, in the following paragraph, the study compares results with similar work. In Section 4 Results and Discussion, the clustering results showed similar visualizations across the three algorithms in classifying SUD for the three substances. Compared to this work, which uses unsupervised machine learning to analyze the data from NSDUH, the majority of studies have used supervised machine learning to classify people with SUD between addicted or not addicted. One such study is, [87] which classified AUD patients and healthy individuals with a classification performance of 89% as the highest accuracy by using logistic regression, which is a supervised machine learning algorithm. Another study is [22] which proposed an application of seven machine learning algorithms

to identify the state of mental health in a target group. In that work, 90% is the highest accuracy of the two best-performing algorithms using both supervised and unsupervised machine learning. In this study, 72.5% is the highest accuracy using silhouette score in the k-means alcohol dataset. The two other previous works use different datasets and different machine learning algorithms. In this sense, supervised machine learning may look to have much higher accuracy. It would be interesting to compare the results with previous studies on SUD that also used unsupervised machine learning.

4.2 Interpretation of Results

In this section, this work discusses the interpretation of the data that is being modeled and analyzed using machine learning. Two types of interpretations are presented: one that is based purely on the machine learning model and another that is based on the interpretation by a mental health professional.

4.2.1 By Machine Learning

In Table 4.3 and 4.4 a summary of the machine learning interpretations and the corresponding interpretations by a mental health professional are presented. The machine learning interpretations are based on the graph visualizations by outlining all the observed trends in the clustering model. The results are compared by regions of the density of the clusters, the range of coverage, and lastly by the boundaries between the clusters. In all the clustering results, two major classifications are observed: a major cluster and a minor cluster. The major cluster is a dense region and the minor cluster is a less dense region. It is interpreted that the major cluster are the behaviors of persons with no significant symptoms of SUD. In this case, they cover significantly large variations of activities and are closely related to each other. This work further interprets this as related to the everyday activities of persons with no SUD which is significantly similar to others with no SUD. The minor cluster is interpreted as the behaviors of persons showing symptoms of SUD. In this case, each person has a unique response to SUD that somehow restricts them from performing normal activities, such that these activities are significantly different from the other persons with symptoms of SUD. This results in a less dense cluster because the person's activities are now limited and are different from others with SUD. The major and the minor clusters of the alcohol dataset are closest to each other compared to the major and minor clusters of marijuana and cocaine. The distance between the cluster centroids is the smallest in the alcohol dataset compared to the other two substances. Thus it can be interpreted that it is easier for a person to move from non-AUD to AUD compared to moving from independence to dependence on marijuana or cocaine, and vice versa. It can be further interpreted that the

Intepretation of Results	
By Machine Learning	By a Mental Health Professional
1. Two major classifications are observed (major cluster, minor cluster) • Observations: ■ Major cluster is a dense region ■ Minor cluster is a less dense region • Interpretations: ■ Major cluster behaviors are closely related to each other ♦ Normal people's behavior ♦ No history of substance use ♦ History of substance use but no significant symptoms of SUD ■ Minor cluster behaviors are not as closely related as major cluster ♦ Behavior of people affected by substance dependence ♦ Symptoms of SUD 2. Major and minor clusters of the alcohol dataset • Observations(compared to the other two substances): ■ The clusters are closest to each other ■ Distance between cluster centroids is smallest • Interpretations(compared to the other two substances): ■ Easiest to move from non-AUD to AUD ■ One can become dependent at the fastest rate ■ One can be free from dependence at the fastest rate 3. Major and minor clusters of the marijuana dataset • Observations(compared to alcohol): ■ The clusters are farther from each other ■ Distance between cluster centroids is more • Interpretations(compared to alcohol): ■ More difficult to move from non-CaUD to CaUD ■ One can become dependent at a slower rate ■ One can be free from dependence at a slower rate ■ Marijuana is more addictive than alcohol	1. Interpretations ■ Major cluster behaviors are closely related to each other ♦ People not meeting criteria for DSM-5 SUD ♦ Not meeting criteria may be due to no history of substance use ♦ Alternatively, the individual may have a history of substance use but do not meet the criteria for SUD. ■ Minor cluster behaviors are not as closely related as major cluster ♦ Individuals with SUD may engage in a variety of behaviors as reported on NSDUH that would align with meeting criteria for a given SUD. 2. Interpretations (compared to the other two substances): ■ Alcohol is a commonly used substance. There is greater engagement in alcohol use behaviors among those who drink. ■ There is greater variability in the behaviors of those who do and do not meet diagnostic criteria for AUD. ■ Individuals may modify patterns of alcohol use more readily (relative to other substances), affecting whether they meet diagnostic criteria for AUD or not. 3. Interpretations (compared to alcohol): ■ A greater proportion of individuals who report using marijuana will meet the criteria for cannabis use disorder relative to those reporting alcohol use and meeting the criteria for AUD. ■ NSDUH assessment demonstrates greater alignment between behaviors and meeting DSM-5 criteria for Cannabis use disorder. ■ Individuals using higher levels of marijuana are more likely to meet DSM-5 criteria for Cannabis use disorder than those using at higher levels.

Table 4.3: Interpretations of results by Machine Learning and by a Mental Health Professional.

rehabilitation efforts needed for the person to become free from AUD require the least compared to CaUD or CoUD. The major and minor clusters of the marijuana dataset are farther from each other compared to the alcohol dataset. The distance between the cluster centroids in the marijuana dataset is more than the distance between the cluster centroids in the alcohol dataset. It can be interpreted that, it is more difficult to move from non-CaUD to CaUD compared to non-AUD to AUD, and vice

Intepretation of Results	
By Machine Learning	By a Mental Health Professional
4. Major and minor clusters of the cocaine dataset • Observations(compared to the other two substances): ■ The clusters are farthest away from each other ■ Distance between cluster centroids is largest • Interpretations(compared to the other two substances): ■ Most difficult to move from non-CoUD to CoUD ■ One can become dependent at the slowest rate ■ One can be free from dependence at the slowest rate ■ Cocaine is the most addictive compared to alcohol and marijuana 5. Range of coverage of major and minor clusters of the alcohol dataset • Observations: ■ Major cluster has a wider coverage ■ Minor cluster has a lesser coverage • Interpretations: ■ Major cluster behaviors have a wider range of activities ♦ Normal people's behavior ♦ Non-users, moderate users, and low-risk users ♦ No significant symptoms of AUD ■ Minor cluster behaviors have a lesser range of activities compared to the major cluster ♦ Behavior of people affected by alcohol dependence ♦ Extreme binge drinkers, heavy drinkers ♦ Symptoms of AUD ■ Boundary of major and minor clusters ♦ Large amounts of consumption but not dependent ♦ Binge drinkers 6. Range of coverage of major and minor clusters of the marijuana and the cocaine datasets • Observations: ■ Major cluster has a wider coverage ■ Minor cluster has a lesser coverage • Interpretations: ■ Major cluster behaviors have a wider range of activities ♦ Normal people's behavior ♦ Non-users and low-risk users ♦ No significant symptoms of CaUD or CoUD ■ Minor cluster behaviors have a lesser range of activities compared to the major cluster ♦ Behavior of people with CaUD or CoUD ♦ Symptoms of CaUD or CoUD	4. Interpretations (compared to the other two substances): ■ Individuals using cocaine are more likely to meet DSM-5 criteria for a Stimulant Use Disorder than those using alcohol or marijuana. ■ Cocaine use is associated with a greater likelihood of experiencing other DSM-5 Stimulant Use Disorder criteria relative to alcohol or marijuana. This may be due to differences in the psychopharmacological properties between the 3 substances. 5. Interpretations ■ Major cluster behaviors have a wider range of activities ♦ Individuals can either not drink or drink alcohol, but may not meet the criteria for DSM-5 AUD. ■ Minor cluster behaviors have a lesser range of activities compared to the major cluster ♦ More extreme alcohol behaviors are more likely to be associated with meeting DSM-5 criteria for AUD ■ Boundary of major and minor clusters ♦ Individuals can consume alcohol to varying degrees, but consumption alone does not determine whether someone meets DSM-5 criteria for AUD. 6. Interpretations ■ Major cluster behaviors have a wider range of activities ♦ Individuals may not be using marijuana or cocaine ♦ Individuals using marijuana, but not use consistent with meeting DSM-5 criteria for Cannabis Use Disorder ♦ Individuals who are using cocaine, but not use consistent with meeting DSM-5 criteria for Stimulant Use Disorder ■ Minor cluster behaviors have a lesser range of activities compared to the major cluster ♦ Those engaging in higher rates of marijuana are more likely to meet DSM-5 criteria for Cannabis Use Disorder. ♦ Similarly, those engaging in higher rates of cocaine are more likely to meet DSM-5 criteria for Stimulant Use Disorder.

Table 4.4: Interpretations of results continued from Table 4.3

versa. It can also be said that an individual can become dependent or independent from marijuana at a slower rate compared to alcohol. Furthermore, the behavior of an individual dependent on marijuana is more distinct from normal behavior compared to alcohol. In the cocaine dataset, the major and the minor clusters are farthest away from each other compared to the alcohol and marijuana datasets. The distance between the cluster centroids is the largest in the cocaine dataset compared to the other two substances. It can, therefore, be interpreted that, moving from non-CoUD to CoUD is the most difficult compared to moving from non-AUD to AUD or non-CaUD to CaUD, and vice versa. Thus, one can become dependent or independent from cocaine at the slowest rate compared to alcohol and marijuana. These interpretations could mean that the rehabilitation efforts needed for the person to become free from SUD will take the biggest efforts in the case of CoUD, followed by CaUD, and lastly AUD. For the alcohol dataset, the major cluster behaviors have a wider coverage compared to the minor cluster. Based on this observation, it can be interpreted that the major cluster has a wider range of everyday activities and consists of normal people's behavior who are non-users, moderate users, or low-risk users of alcohol. It can be concluded that these individuals have no significant symptoms of AUD because their everyday behavior is not affected by alcohol. The minor cluster behaviors have a lesser range of coverage compared to the major cluster. It consists of people affected by alcohol dependence resulting in a lesser of range everyday activities. These individuals include extreme binge drinkers or heavy drinkers and are projecting symptoms of AUD because their behavior is limited due to the effects of alcohol. At the boundary between the clusters, lie the behavior of people that can swing from independence to dependence of alcohol, and vice versa. The behaviors at the boundary of the major and the minor clusters consist of individuals who may have large alcohol consumption but are not considered dependent or are considered not very dependent. This drinking pattern is normally found in binge drinkers. In the marijuana and cocaine datasets, there is a wider range of coverage in the major clusters and a smaller range of coverage in the minor clusters. It can thus, be interpreted that, the major clusters have a wider range of activities that consist of normal people's behavior. These individuals may be regarded as non-users or low-risk users of marijuana or cocaine and have no significant symptoms of CaUD or CoUD. The minor clusters have a less range of activities and can be interpreted as the behaviors of people affected by marijuana or cocaine dependence, who are showing symptoms of CaUD or CoUD such that their everyday activities are constrained within a lesser range. Comparing all three substances, it can be observed that the region for CaUD covers the widest range, followed by AUD, and lastly CoUD. It can be said that the behaviors of people with CaUD are the least restrictive such that, their everyday activities are the least constrained. This makes individuals with CaUD more functional and their behavior is closer to people with no SUD, as compared to people with AUD or CoUD. Thus the manifestation

of SUD is least obvious with CaUD, followed by AUD, and the most obvious manifestation of SUD is CoUD. The results complement the existing research that risk perceptions related to the use of marijuana are becoming more non-restrictive across all ages, especially in adults, emerging adults, and middle adults [88]. In addition, the distance between the cluster boundaries in the alcohol dataset is the least, followed by marijuana, and lastly cocaine. It can be said that the behaviors of the people at the cluster boundaries of AUD and non-AUD are similar to each other. Thus, it is difficult to distinguish AUD and non-AUD behaviors at the boundaries. The distance at the cluster boundaries of the marijuana dataset is farther compared to the alcohol. This means that the behaviors of people at the cluster boundaries of CaUD and non-CaUD are more distinct compared to alcohol. Thus, the difference in the behaviors of CaUD and non-CaUD is observable. And lastly, the distance at the cluster boundaries of the cocaine dataset is farthest compared to alcohol and marijuana. This means that the behaviors of people at the cluster boundaries of CoUD and non-CoUD are the most distinct compared to alcohol and marijuana. Thus, the difference in behaviors is most observable between CoUD and non-CoUD compared to the two other substances. It was concluded that in the DBSCAN results for the alcohol dataset, the smallest cluster represents respondents with severe symptoms of AUD who can be regarded as heavy drinkers, and for the DBSCAN results in the marijuana, and cocaine datasets, participants are projecting heterogeneous symptoms for potential poly-substance use disorder. This is because a study by [89] suggests marijuana dependence and cocaine dependence are related to the increased likelihood of binge drinking. It is also possible that the model has detected potential endophenotypes for CaUD and CoUD as according to [90] both of these drugs are heritable by at least 50%. Lastly, a study by [25] interprets features based on non-normative socialization SUD emphasis to manifest from psychological and health issues to social maladjustment. That work further explains that the strongest indicators of SUD risk are swearing, poor behavior, and irritability. The authors interpret that it could result in affective dysregulation and social interaction problems which could also result in irregular circadian rhythms. In terms of SUD prediction, it is mentioned that phenotypic characteristics and environmental factors ranked among the top two features which machine learning combined to facilitate the prioritization of prevention tactics. Compared to this works's interpretations from observed machine learning results, these interpretations need formal training in psychology. Our interpretation is based on data visualizations provided by unsupervised machine learning, which do not need any formal background in psychology. Thus, validation from a mental health professional is needed. The statistical results from the NSDUH reported that 5.3% of the total population aged 12 years or older had AUD, about 1.8% of individuals from the total population were diagnosed with CaUD, and 0.4% of the respondents had CoUD. It was also confirmed that 7.4% of people aged 12 years or older have developed SUD for alcohol, marijuana, and cocaine.

Of the 7.4% diagnosed with SUD, about 71.1% of people had AUD, 23.5% were diagnosed with CaUD, and about 4.9% had CoUD [45].

4.2.2 By a Mental Health Professional

Machine learning results point to different patterns in the classifications of SUDs when examining alcohol, marijuana, and cocaine in the NSDUH data. For alcohol and marijuana, the boundaries between the group with and without AUD and CaUD, respectively, were less distinct. This pattern may reflect that DSM-5 diagnostic criteria focus on symptoms (e.g., tolerance, withdrawal, and negative consequences of substance use), rather than specific substance use behaviors (e.g., frequency of use) that are assessed by NSDUH. In the case of alcohol and marijuana, results may be driven by the more widespread use of both substances and that alcohol or marijuana use may not correspond with meeting DSM-5 diagnostic criteria. In contrast, there was a clear separation between those who were and were not classified as addicted to cocaine. This may be due to the lower frequency of cocaine use and the greater probability that individuals using cocaine may meet diagnostic criteria for a Stimulant Use Disorder. When examining the spread of the data across the three substances, different patterns emerged by substance. For alcohol and marijuana, there was a greater spread in the results for the non-addicted groups, pointing to a greater range of behaviors represented relative to the cocaine group. This is consistent with patterns of use for these substances in the U.S. Additionally, there was greater density for the alcohol and marijuana groups identified as addicted relative to cocaine. This likely reflects a more limited range of behaviors endorsed for alcohol and marijuana relative to cocaine.

4.3 Brain-computer interface experiments using alcohol stimulus

In this section, the study presents three main types of results derived from the processing of the recorded data. This work starts with an analysis of the power spectrum of the theta, alpha, and beta frequency rhythms for low-risk consumption, harmful consumption, and the likelihood of AUD. It then discusss the transitions of the brain signals during the first and last phase of alcohol intake. Finally, the overall cluster results for all the participants combined are described, by showing the brain activity of the subjects with hazardous alcohol use, dependence symptoms, and lastly those with severe alcohol use.

4.3.1 Power spectral density analysis

In the following paragraphs, the study presents the PSD graphs for participants from each category of low-risk consumption, harmful or hazardous consumption, and the likelihood of AUD that is based on the responses of the AUDIT. It discusses the theta, alpha, and beta frequency bands' results during the first and last phase of alcohol intake for three subjects from each category. All the PSD graphs for all 17 participants are shown as supplementary material. Lastly, it will discuss the general observations and interpretations of the results based on the three subjects from each category.

4.3.1.1 Theta band

Theta band	
Observations	Interpretations
1. Low-risk consumption • The PSD ranges from 0.005 to 0.065 $\mu V^2/Hz$ in epoch 1 and from 0.005 to 0.055 $\mu V^2/Hz$ in epoch 6 • A higher degree of signal overlap compared to the harmful consumption and likelihood of AUD in both epochs • Lowest range of theta power compared to the harmful consumption and likelihood of AUD in epochs 1 and 6 2. Harmful or hazardous consumption • The PSD ranges from 0.005 to 0.09 $\mu V^2/Hz$ in epoch 1 and from 0.005 to 0.15 $\mu V^2/Hz$ in epoch 6 • The degree of signal overlap is lesser in both epochs compared to the low-risk consumption • The range of theta power is more compared to the low-risk consumption for both epochs 1 and 6 3. Likelihood of AUD • The PSD ranges from 0.00 to 0.10 $\mu V^2/Hz$ in epoch 1 and from 0.01 to 0.28 $\mu V^2/Hz$ • For both epochs, the degree of signal overlap is lesser compared to the low-risk and harmful consumption domains • The range of theta power is the largest compared to the low-risk and hazardous risk domain for both epochs 1 and 6	1. An increase in alcohol consumption causes an increase in theta spectral power as the amount of alcohol increases. 2. A higher degree of signal entanglement results from an activation of the inhibitory and motor responses caused by alcohol consumption. This is because these are regulated by the theta band [91]. 3. signal entanglement could also indicate acute changes in brain activity in response to alcohol intoxication. This is because we can observe mild intoxication effects in most people after exactly two standard drinks [14]. 4. A higher frequency of alcohol consumption results in higher theta energy. This is evident in that people with low-risk consumption have the lowest range of theta power.

Table 4.5: Observations and the Interpretations of the PSD results obtained from the theta frequency rhythm.

This frequency rhythm lies within the range (4-8Hz) and is mostly responsible for the regulation of learning, inhibitory control processes, and intuition. It is generated across all parts of the cortex

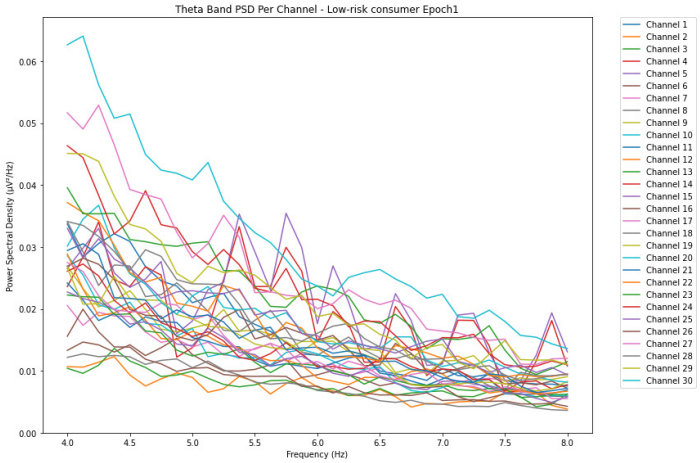


Figure 4.16: Epoch 1 represents the initial phase of alcohol intake for the low-risk consumption in the theta band. We can observe a range of 0.065 μV²/Hz PSD and a high degree of signal entanglement.

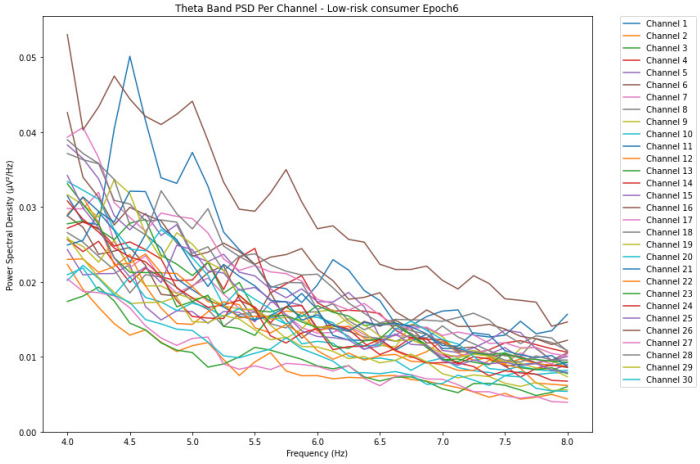


Figure 4.17: Epoch 6 represents the last phase of alcohol intake for the low-risk consumption in the theta band. The average PSD range was reduced by 0.01 μV²/Hz. The degree of signal entanglement is less diverse compared to epoch 1.

and acts as a repository for memories and emotions [91].

4.3.1.1.1 Observations It can be observed that the PSD of a low-risk consumer ranges from 0.005 to 0.065 μV²/Hz in Fig. 4.16 and from 0.005 to 0.055 μV²/Hz in Fig. 4.17. The PSD of the harmful or hazardous consumer ranges from 0.005 to 0.09 μV²/Hz in Fig. 4.18 and from 0.005 to 0.15 μV²/Hz in Fig. 4.19. Finally, the PSD of a participant with the likelihood of AUD ranges from 0.00 to 0.10 μV²/Hz in Fig. 4.20 and from 0.001 to 0.28 μV²/Hz in Fig. 4.21. The participant with low-risk consumption has a higher degree of signal entanglement in epoch 1, followed by the one with harmful consumption, and lastly the subject with the likelihood of AUD. On the contrary, in epoch 6, there is a higher degree of signal entanglement for harmful or hazardous risk consumer, followed by the one with low-risk consumption, and lastly the subject with the likelihood of AUD. Based on these results, during the first phase of alcohol intake, the participant with the likelihood

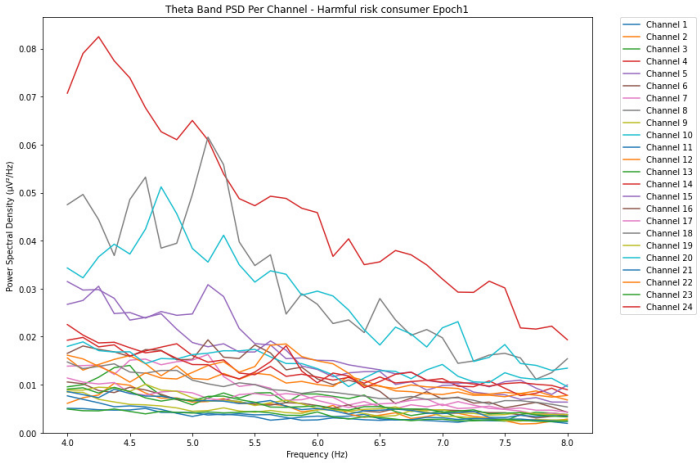


Figure 4.18: Epoch 1 represents the initial phase of alcohol intake for the harmful or hazardous consumption in the theta band. We can observe a range of 0.08 $\mu V^2/Hz$ PSD and a lesser degree of signal entanglement compared to the low-risk consumption.

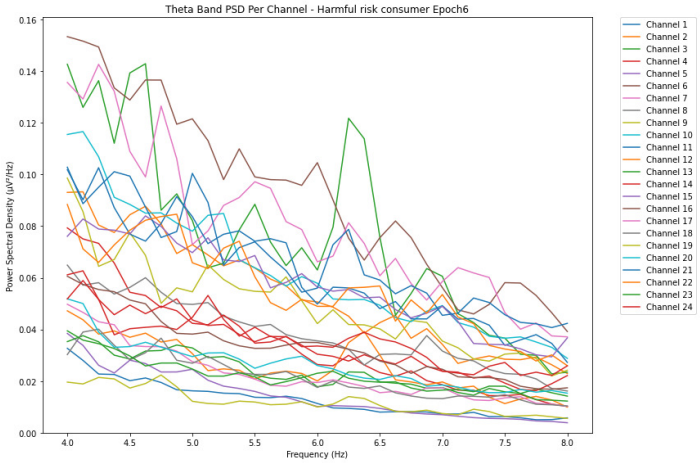


Figure 4.19: Epoch 6 represents the last phase of alcohol intake for the harmful or hazardous consumption in the theta band. The average PSD range increased by 0.065 $\mu V^2/Hz$. The degree of signal entanglement is less diverse compared to epoch 1.

of AUD has the highest range of theta power, followed by the harmful or hazardous risk consumer, and lastly, the low-risk consumer. Similarly, during the last phase of alcohol intake, the subject with the likelihood of AUD has the highest range of theta power, followed by harmful or hazardous risk consumer and lastly low-risk consumer.

4.3.1.1.2 Interpretations In general, for the theta band, more alcohol consumption results in more spectral power. This is true for all types of alcohol consumption. However, for people with a lesser frequency of consumption, the level of activation is much higher compared to the highest frequency of consumption. For the data gathered in this experiment, the highest frequency of consumption has five times more spectral power than the lowest frequency of consumption. This means that people with higher alcohol consumption are calmer and therefore more coherent in their level

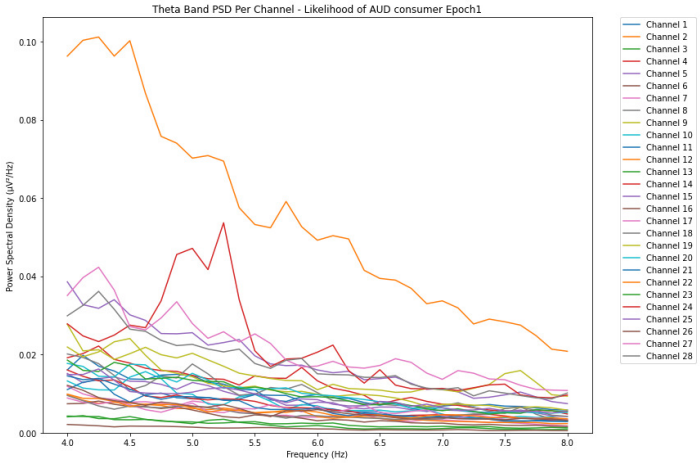


Figure 4.20: Epoch 1 represents the initial phase of alcohol intake for the likelihood of AUD in the theta band. We can observe a range of 0.10 μV²/Hz PSD and the degree of signal entanglement is less and the lower set of signals shows more stability.

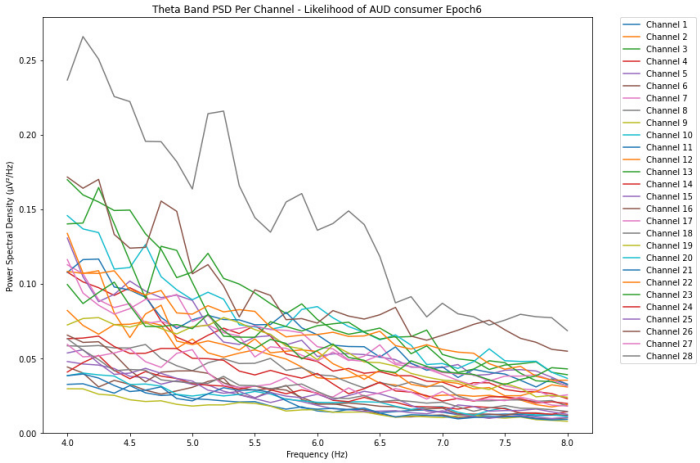


Figure 4.21: Epoch 6 represents the last phase of alcohol intake for the likelihood of AUD in the theta band. The average PSD range increased by 0.14 μV²/Hz between epochs 1 and 6. The degree of signal entanglement is less diverse in both epochs compared to the low-risk and harmful consumption.

of concentration thus they tend to be more functional in performing normal tasks compared to those with less consumption. The high degree of signal entanglement results from the activation of inhibitory and motor responses caused by alcohol consumption as these are regulated by the theta band [91]. The level of activation is much higher for people with lesser frequency of use because they project more acute performance impairment symptoms due to higher alcohol consumption. In most people, mild intoxication can be observed after two standard drinks [14]. Table 4.5 summarizes the results of the theta band.

4.3.1.2 Alpha band

The alpha band is a state of the brain that predominates through the resting state and enhances an individual's relaxation of the mental state. Its frequency ranges between 8-12Hz and it is a source for

mental coordination and consciousness that does not involve any cognitive task. Its power spectrum has been proven to be mostly suppressed during body activities with eyes wide open [92].

Alpha band	
Observations	Interpretations
1. Low-risk consumption • The PSD ranges from 0.002 to 0.014 μV^2/Hz in epoch 1 and from 0.002 to 0.0158 μV^2/Hz in epoch 6 • Similar to the theta band, we can observe a higher degree of signal overlap in epochs 1 and 6 compared to the harmful consumption and likelihood of AUD • Lowest range of alpha power compared to the harmful consumption and likelihood of AUD in epochs 1 and 6 • Compared to the theta band, we can observe a decrease in the average PSD for the low-risk, harmful risk, and likelihood of AUD 2. Harmful or hazardous consumption • The PSD ranges from 0.001 to 0.024 μV^2/Hz in epoch 1 and from 0.002 to 0.048 μV^2/Hz in epoch 6 • The degree of signal overlap is lesser in both epochs compared to the low-risk consumption • The range of alpha power is more compared to the low-risk consumption for both epochs 1 and 6 3. Likelihood of AUD • The PSD ranges from 0.000 to 0.0225 μV^2/Hz in epoch 1 and from 0.005 to 0.075 μV^2/Hz • For both epochs, the degree of signal overlap is lesser compared to the low-risk and harmful consumption domains • The range of alpha power is the largest compared to the low-risk and hazardous risk domain for both epochs 1 and 6	1. Similar to the theta band, more alcohol consumption results in more spectral power. This is because the highest range of alpha power was observed in epoch 6 compared to epoch 1. 2. Similarly, the frequency of use of alcohol results in more spectral power. This has been shown by a higher average PSD for the individual with the likelihood of AUD as compared to those with low-risk and harmful risk consumption. 3. The alpha rhythm predominates in the resting state thus its spectral power is suppressed during body activities with eyes wide open [92]. This is why it has the lowest range of PSD compared to the theta and beta bands. 4. The alpha band promotes mental coordination and enhances overall relaxation. The increase in the average power spectrum is caused by cortical activations and alert mechanisms from increasing alcohol amount [93].

Table 4.6: Observations and the Interpretations of the PSD results obtained from the alpha frequency rhythm.

4.3.1.2.1 Observations Compared to the results of the theta band, there is a significant decrease of alpha in the range of the PSD for all the subjects in all the epochs. Although the results of the alpha band are lesser than theta, there is an increase in the average PSD between epochs 1 and 6 for all three types of consumption. It can be observed that the PSD of low-risk consumer ranges from 0.002 to 0.014 μV^2/Hz in Fig. 4.22 and from 0.002 to 0.0158 μV^2/Hz in Fig. 4.23.

The PSD for harmful risk consumer in Fig. 4.24 ranges from 0.002 to 0.024 μV^2/Hz and from

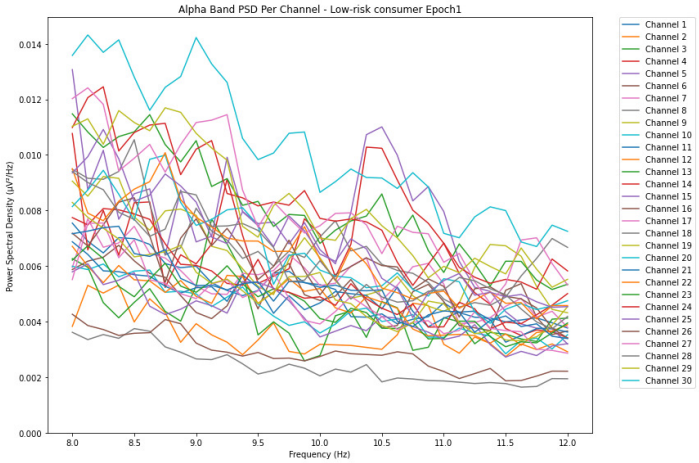


Figure 4.22: Epoch 1 represents the initial phase of alcohol intake for the low-risk consumption in the alpha band. We can observe a range of 0.014 μV²/Hz PSD and a higher degree of signal entanglement.

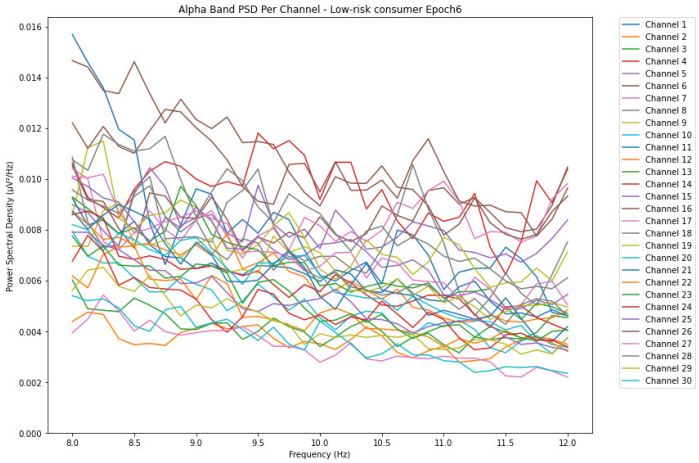


Figure 4.23: Epoch 6 represents the last phase of alcohol intake for the low-risk consumption in the alpha band. The average PSD range was increased by 0.002 μV²/Hz. The degree of neural overlap is almost similar to epoch 1.

0.005 to 0.48 μV²/Hz in Fig. 4.25. Lastly, the PSD for an individual with the likelihood of AUD in Fig. 4.26 ranges from approximately 0.000 to 0.0225 μV²/Hz in epoch 1 and from 0.005 to 0.075 μV²/Hz in Fig. 4.27. Similar to the theta band, there is a higher degree of signal entanglement for the subject with low-risk consumption, followed by the harmful or hazardous risk consumer, and lastly for the subject with the likelihood of AUD in epoch 1. For epoch 6, it can be observed that a higher degree of signal entanglement in the subject with harmful or hazardous risk consumption, followed by the low-risk consumer, and lastly one with the likelihood of AUD.

4.3.1.2.2 Interpretations In the alpha, the same trend of activation is observed as in the theta band. However, the value of the spectral power is lesser because it is mostly suppressed during body activities with eyes wide open. A higher alpha power is more prominent at the resting state and thus more alcohol consumption results in less relaxation and more agitation [92]. People with lesser

alcohol consumption can be easily disturbed by outside stimuli compared with people with higher consumption as they can be calmer. Increasing the amount of alcohol consumption results in cortical activations and alert mechanisms and thus causes less relaxation [93]. The summary of observations and interpretations of the alpha band are summarized in table 4.6

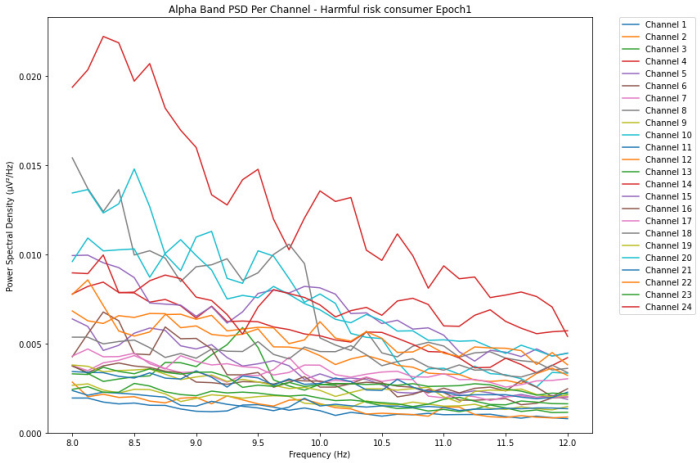


Figure 4.24: Epoch 1 represents the initial phase of alcohol intake for the harmful or hazardous consumption in the alpha band. We can observe a range of 0.024 μV²/Hz PSD and a lesser degree of signal entanglement compared to the low-risk consumption.

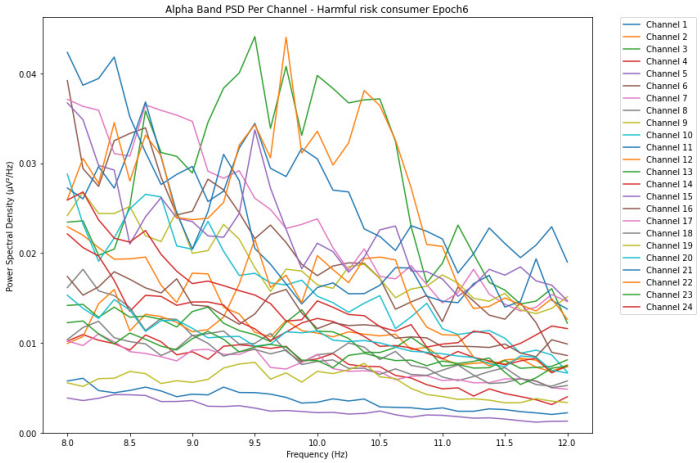


Figure 4.25: Epoch 6 represents the last phase of alcohol intake for the harmful or hazardous consumption in the alpha band. The average PSD range increased by 0.024 μV²/Hz between epochs 1 and 6. The degree of signal entanglement is more compared to epoch 1.

4.3.1.3 Beta band

In the beta rhythm, the frequency ranges from 12-35Hz which can further be subdivided into low beta (12-15Hz), mid beta (15-18Hz), and high beta (above 18Hz). It is a state of the brain in which sensory-motor control tasks are processed and occur during information processing, decision-making, and problem-solving. It is recorded in the frontal regions of the brain [94].

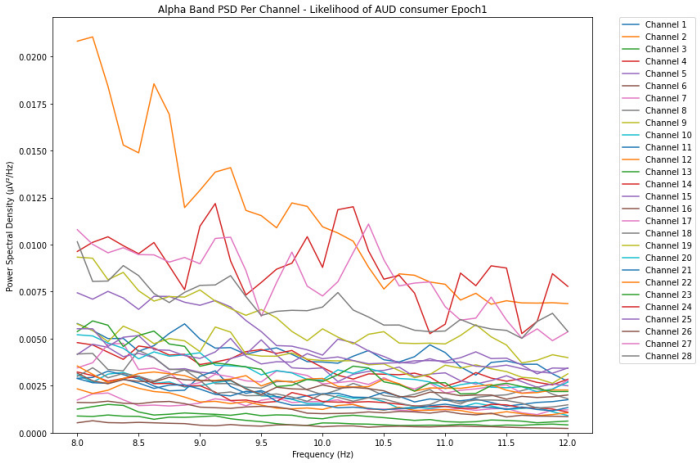


Figure 4.26: Epoch 1 representing the initial phase of alcohol intake for the likelihood of AUD in the alpha band. We can observe a range of 0.023 μV²/Hz PSD and a lesser degree of signal entanglement compared to the low-risk and harmful consumption. The lower set of signals have a smooth and stable trend while the higher signals are more jumpy.

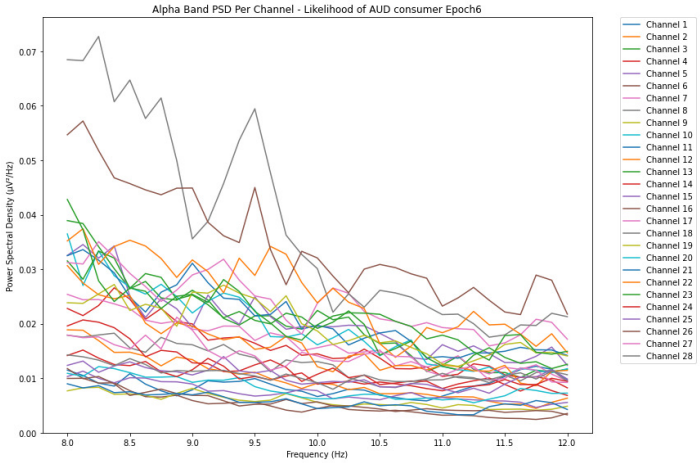


Figure 4.27: Epoch 6 representing the last phase of alcohol intake for the likelihood of AUD in the theta band. The average PSD range increased by 0.053 μV²/Hz between epochs 1 and 6. The degree of neural entanglement is less diverse in both epochs compared to the low-risk and harmful consumption.

4.3.1.3.1 Observations The PSD of low-risk consumer ranges from 0.0005 to 0.01 μV²/Hz in Fig. 4.28 and from 0.000 to 0.045 μV²/Hz in Fig. 4.29. For harmful or hazardous risk consumer in Fig. 4.30, the PSD ranges from approximately 0.000 to 0.0065 μV²/Hz and approximately 0.000 to 0.02 μV²/Hz in Fig. 4.31. Lastly, the PSD for the subject with the likelihood of AUD in Fig. 4.32 ranges from 0.000 to 0.013 μV²/Hz and from 0.000 to 0.024 μV²/Hz in Fig. 4.33.

The signals appear to fluctuate independently but the degree of signal entanglement is more significant in the low-risk consumer, followed by harmful or hazardous risk consumer, and lastly the subject with the likelihood of AUD. It can be observed that some signals have peaks at certain frequencies, with the low-risk consumer having more pronounced peaks, followed by harmful risk consumer, and lastly, the subject with the likelihood of AUD in epoch 1. In epoch 6 of the low-risk

Beta band	
Observations	Interpretations
1. Low-risk consumption • The PSD ranges from 0.0005 to 0.01 $\mu V^2/Hz$ in epoch 1 and from 0.000 to 0.045 $\mu V^2/Hz$ in epoch 6 • The signals fluctuate independently however the degree of neural overlap is more significant in the low-risk consumption compared to harmful risk and likelihood of AUD • The low-risk consumption has more pronounced peaks with the most significant observed at 15Hz • The lowest range of beta power can be observed compared to the other two domains 2. Harmful or hazardous consumption • The PSD ranges from 0.000 to 0.0065 $\mu V^2/Hz$ in epoch 1 and from 0.000 to 0.02 $\mu V^2/Hz$ in epoch 6 • The range of beta power is lesser compared to the low-risk consumption for both epochs 1 and 6 3. Likelihood of AUD • The PSD ranges from 0.000 to 0.013 $\mu V^2/Hz$ in epoch 1 and from 0.000 to 0.024 $\mu V^2/Hz$ • For both epochs, the degree of signal overlap is lesser compared to the low-risk and harmful consumption domains • The range of beta power is lesser compared to the low-risk but more compared to the harmful risk consumption for both epochs 1 and 6	1. For low-risk consumption, the spectral power is more pronounced in the sensory-motor rhythm (12-15Hz). It can reflect an imbalance between the excitatory and inhibitory neural responses due to increasing alcohol amount for an individual whose consumption is low [95]. 2. A significant increase in the beta spectral power is normally produced before developing an AUD [96]. In our results, PSD in the beta was lower than the theta and alpha results. This could indicate that the participants project AUD symptoms when the amount of alcohol consumption increases but are not necessarily suffering from AUD 3. Additionally, the increase in the beta power does not entirely depend on the rate of consumption of alcohol, and as such we can reduce the possibility of these participants being alcohol dependent. This is because an increase in the beta spectral power can be related to a family history of AUD [17]. This further supports our argument that these participants may be prone to dependence if they increase their rate and frequency of alcohol consumption. 4. Pronounced peaks could potentially indicate personality traits that are associated with the risk of developing AUD like a family history of alcoholism. This is because results in beta have been considered a trait marker in AUDs [96].

Table 4.7: Observations and the Interpretations of the PSD results obtained from the beta frequency rhythm.

consumer, there is a significant peak at 15Hz.

4.3.1.3.2 Interpretations Similar to the theta and alpha bands, the beta has more spectral power for a higher frequency of alcohol compared to lesser alcohol consumption. This is because beta is about focus and concentration thus the person with the least alcohol consumption has less motor control for the same amount of consumed alcohol compared to the person with the highest frequency of consumption. The results are summarized in table 4.7.

The spectral power of the low-risk consumer is more pronounced and significant in the sensory-motor rhythm (12-15Hz). This rise in beta power may indicate an imbalance between the excitatory and inhibitory neural responses resulting from increasing alcohol amount for an individual whose consumption is low [95]. A study by [96] suggests that an increase in beta spectral power is normally produced before developing AUD. In the results of this study, the PSD in the beta band was

generally lower than the theta and alpha rhythms. This may be evidence of AUD vulnerability when the frequency of consumption is increased. It can be said that the subjects are projecting symptoms of AUD due to the increasing alcohol amount but do not necessarily suffer from AUD. Additionally, the increase in the beta spectral power does not entirely depend on the rate and frequency of use of alcohol. As such, the possibility of the subjects being alcohol dependent is reduced because the increase of the beta power may be related to a family history of AUD [17]. This further supports our argument that the subjects in the study may be prone to AUD if they increase their frequency of alcohol consumption. Lastly, the pronounced peaks observed in the beta results may indicate personality traits associated with the risk of developing AUD because results in beta have been considered a trait marker according to [96]. This is to say that the beta spectral power can be used to diagnose potential endophenotypes that put individuals at risk of developing AUD before the presence of symptoms.

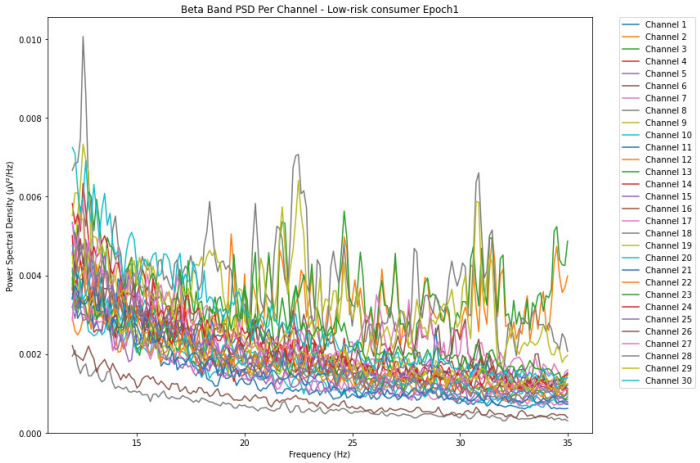


Figure 4.28: Epoch 1 represents the initial phase of alcohol for the low-risk consumption in the beta band. We can observe a range of 0.0095 μV^2/Hz and a higher degree of spiky signals.

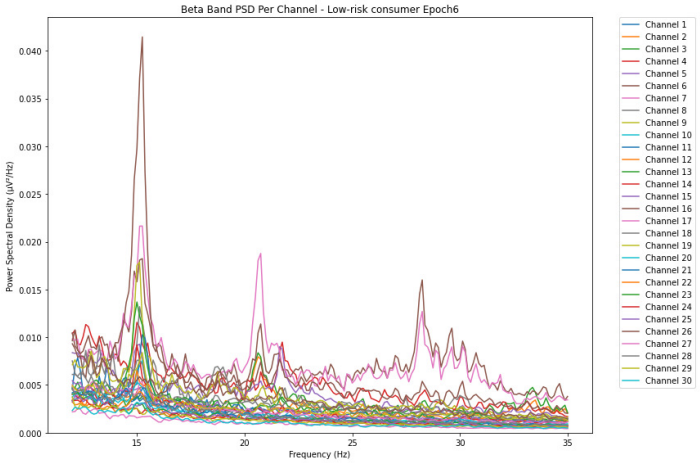


Figure 4.29: Epoch 6 represents the last phase of alcohol intake for the low-risk consumption in the beta band. The average PSD increased by 0.035 μV^2/Hz with a there is a significant spike at 15Hz.

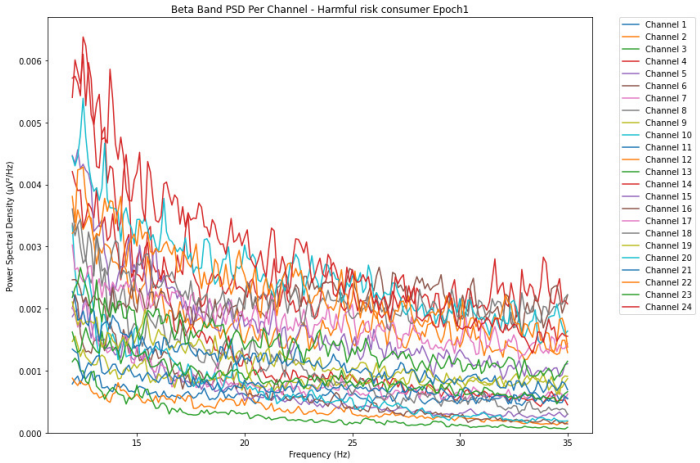


Figure 4.30: Epoch 1 represents the initial phase of alcohol for the harmful or hazardous risk consumption in the beta band. We can observe a range of 0.065 μV²/Hz and a lower degree of signal spikes compared to the low-risk consumption. We can observe a decrease in the PSD of these signals as the frequency increases.

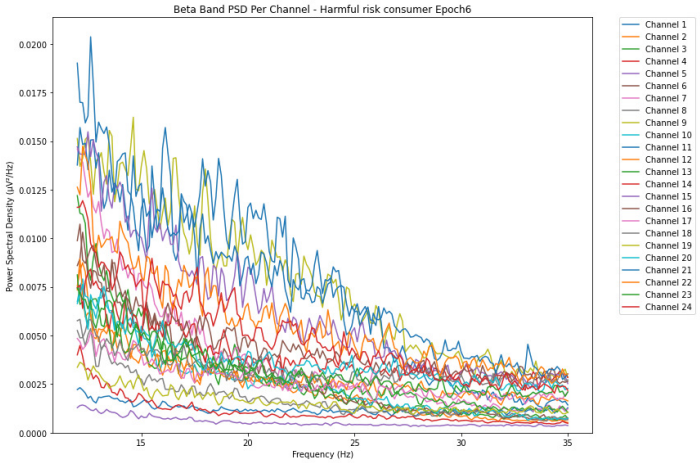


Figure 4.31: Epoch 6 represents the last phase of alcohol intake for the harmful or hazardous risk consumption in the beta band. The average PSD increased by 0.0135 μV²/Hz with a lesser degree of signal spikes. Similar to epoch 1, a decrease in the PSD of these signals as the frequency increases is observed.

In general, the person with a lesser frequency of alcohol consumption is greatly affected by the effects of alcohol for the same amount of alcohol consumed compared to the person with a higher frequency of consumption. This can hinder their ability to make proper judgments and informed decision-making, leading to impaired motor coordination [97]. Results in theta have the highest PSD across all the subjects compared to the other two bands because a study by [96] suggests that theta is considered a state marker in alcoholism. This means that the spectral power of the theta band can be used as a biomarker to diagnose the presence of AUD. According to [17], the spectral power in EEG can be used to examine the extent of alcohol use by showing neural changes associated with its consumption. The results in theta, alpha, and beta bands can correlate with the severity of AUD and how it affects the brain.

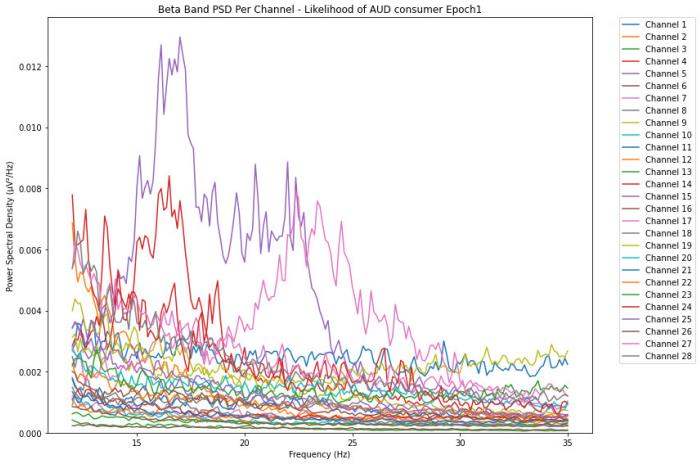


Figure 4.32: Epoch 1 represents the initial phase of alcohol intake for the likelihood of AUD in the beta band. We can observe a PSD of 0.013 $\mu V^2/Hz$ and more stability in the lower set of signals. The signals spikes are less significant and the value of the PSD increases as the frequency decreases.

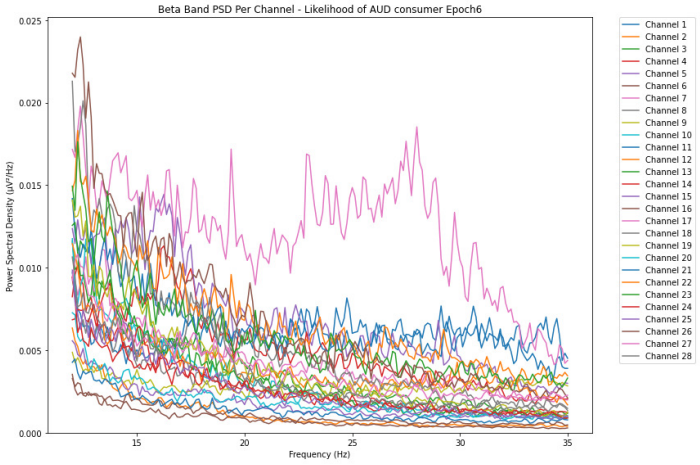


Figure 4.33: Epoch 6 represents the last phase of alcohol intake for the likelihood of AUD in the beta band. The average PSD increased by 0.011 $\mu V^2/Hz$ and there is more signal fluctuations compared to epoch 1. Similar to the harmful or hazardous risk consumption, the power spectrum reduces as the frequency increases.

4.3.2 Transition based clustering results

In the following paragraphs, the study discusses the transitions of the brain signals from the initial to the final phase of alcohol intake for low-risk consumption, harmful or hazardous consumption, and the likelihood of AUD. These transitions will be used to understand an individual's cognitive and behavioral changes with increasing alcohol amount. The results of epoch 1 represent the initial phase of alcohol intake while those for epoch 6 represent the last phase of alcohol intake. The observations and interpretations are discussed from epoch 1 followed by epoch 6 and summarized in table 4.8.

4.3.2.0.1 Observations In Fig. 4.34, the clustering results indicate the three major clusters that represent the brain activity of participants at the first instance of alcohol intake. More dispersed data points are observed for the green cluster, followed by the red cluster, and lastly green red cluster. The

Transition based clustering results	
Observations	Interpretations
1. Epoch 1 (Initial phase of alcohol intake) • Three major clusters representing the results in the first instance of alcohol intake • More dispersed data points in the green cluster, followed by the red cluster, and lastly the blue cluster • The red cluster is denser than the blue and green clusters and covers a wider range • The data points in the red cluster fall between the blue and the green clusters 2. Epoch 6 (Final phase of alcohol intake) • Three major clusters in the last instance of alcohol intake • The data points are more confined and less dispersed compared to the ones in epoch1 • The data points in the blue and green cluster intersect towards the larger values of the PCA1 axis • Some data points in the green cluster overshoot towards the larger values of the PCA1 axis	1. Epoch 1 (Initial phase of alcohol intake) • The three major clusters represent the brain activity of people with low-risk consumption, harmful or hazardous risk consumption, and the likelihood of AUD • The blue cluster represents subjects with the likelihood of AUD, the red cluster represents the brain activity of people with harmful or hazardous consumption while the green represents those with low-risk consumption • The dispersion of data points indicates the dynamic neurophysiological changes and the activation of brain activity induced by alcohol consumption [96] • Similarly, individuals experience more brain activation in lesser alcohol consumption, followed by hazardous or harmful alcohol consumption, and lastly those with the likelihood of AUD 2. Epoch 6 (Final phase of alcohol intake) • The brain activity of people with low-risk consumption and the likelihood of AUD intersect because according to [98] these types of consumers may project similar characteristics during binge drinking episodes. • Data overshoot indicates the high spectral power that is mostly associated with the activation of the inhibitory and motor responses caused by increasing the amount of alcohol for people with low consumption. This is because an increase in alcohol consumption causes an increase in the spectral power of the theta, alpha, and beta frequency bands • The brain activity of people that have a moderate to high alcohol consumption project less motor control impairment compared to those with low-risk consumption [95] and thus are calmer for the same amount of alcohol consumption

Table 4.8: Observations and the Interpretations of the transition-based clustering results.

red cluster is dense than the blue and green clusters and covers a wider range than the other two. The data points in the red cluster fall between the blue and the green clusters.

In Fig. 4.35, three major clusters can also be observed that represent the brain signals of participants at the last instance of alcohol intake. Compared to the epoch 1 cluster results, the data points are more confined and less dispersed. The data points in the blue and green clusters intersect towards the larger values of the PCA1 axis. Similarly, some data points in the green cluster overshoot towards the larger values of the PCA1 axis.

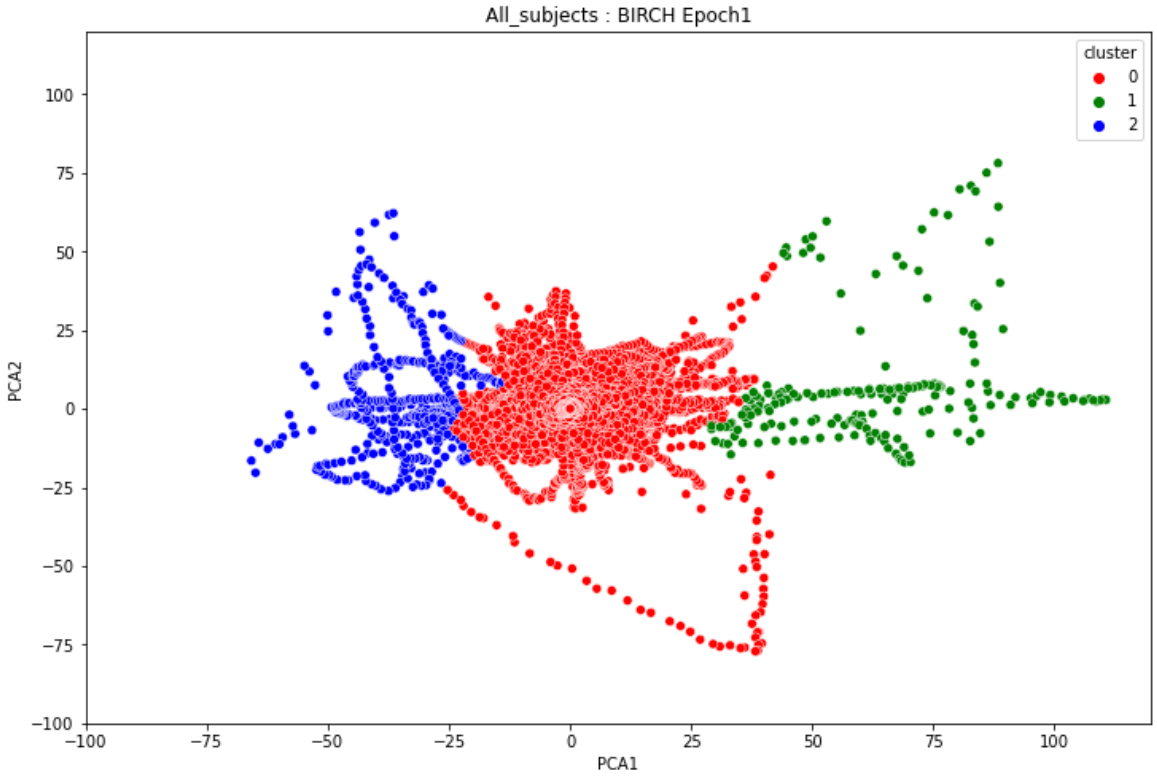


Figure 4.34: The clustering results in the first instance of alcohol consumption for all 17 participants. The green cluster has more dispersed data points compared to the red and blue clusters.

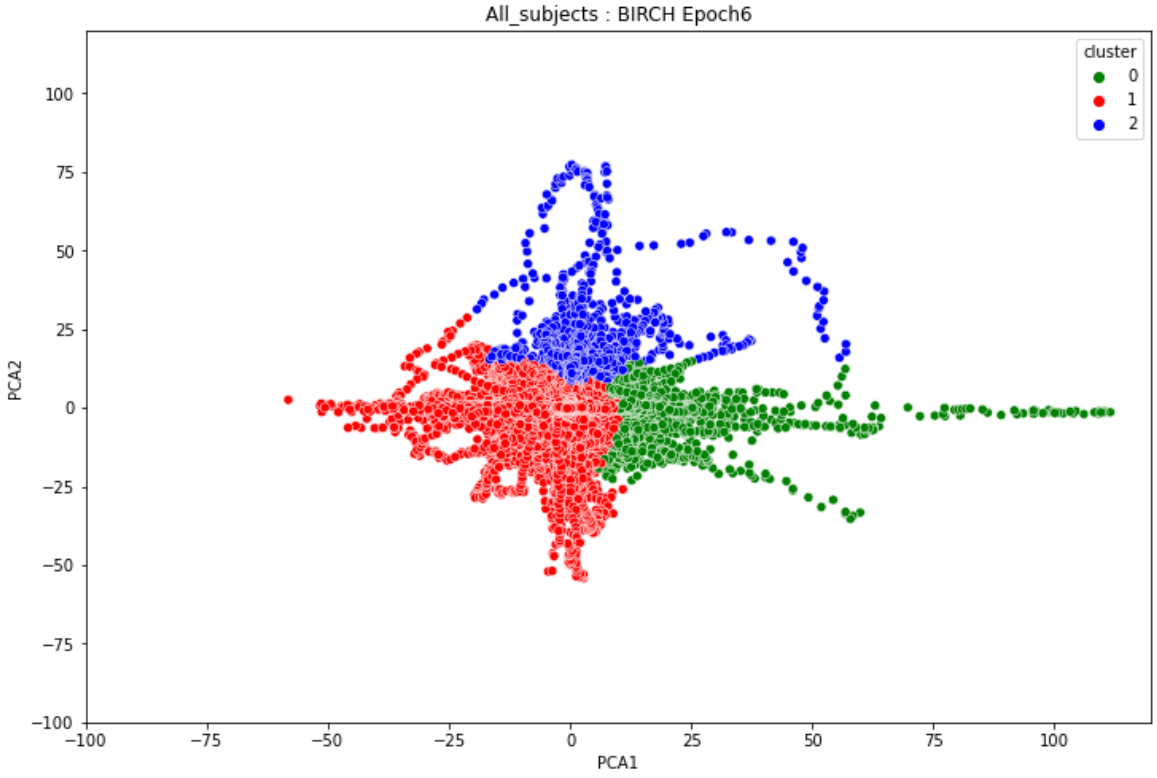


Figure 4.35: The clustering results in the last instance of alcohol consumption for all 17 participants. In this graph, the data points are more confined and have a lesser dispersion.

4.3.2.0.2 Interpretations The three major clusters in Fig. 4.34 and Fig. 4.35 represent the brain activity of people with low-risk alcohol consumption, hazardous or harmful alcohol consumption, and the likelihood of AUD. The blue cluster in Fig. 4.34 represent the brain activity of people with the likelihood of AUD, the red cluster represents those with harmful or hazardous consumption and lastly the green cluster represents those with low-risk consumption. The dispersion of the data points in Fig. 4.34 indicates the neurophysiological changes and the brain activation induced by alcohol consumption. In addition, individuals experience more brain activation and a higher degree of neural entanglement for lesser alcohol consumption compared to higher alcohol consumption. The brain activity of people with low-risk consumption and the likelihood of AUD in Fig. 4.35 intersect because according to [98], these consumers may project similar characteristics during binge drinking. The data overshoot in Fig. 4.35 of the green cluster indicates the higher spectral power that is mostly associated with the activation of the inhibitory and motor responses caused by increasing amount of alcohol for individuals with low consumption. An increase in alcohol consumption causes an increase in the spectral power of the theta, alpha, and beta frequency bands. Lastly, the brain activity of people who have moderate to high alcohol consumption projects less motor control impairment compared to those with low-risk consumption [98] and thus are calmer for the same amount of alcohol consumption.

4.3.3 Overall clustering results

Domains	Questions	Item Content	Responses
Hazardous Alcohol use	1	**Frequency of drinking**	31
	2	**Typical quantity**	45
	3	**Frequency of heavy drinking**	38
Dependence symptoms	4	**Impaired control over drinking**	21
	5	**Increased salience of drinking**	13
	6	**Morning drinking**	25
Severe Alcohol use	7	**Guilt after drinking**	28
	8	**Blackouts**	28
	9	**Alcohol-related injuries**	9
	10	**Others concerned about drinking**	14

Table 4.9: The three domains of the AUDIT with the total responses recorded from the 17 participants in the alcoholism experiments.

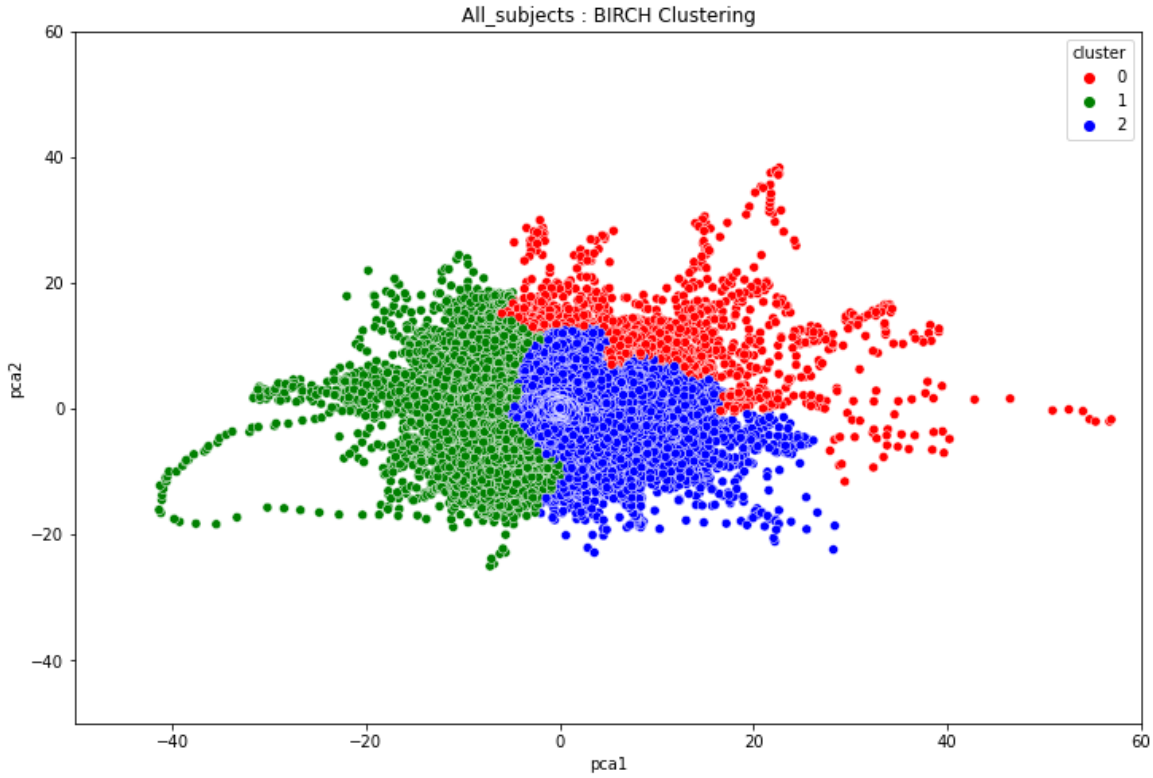


Figure 4.36: The overall clustering results of all the 17 participants from the first phase to the last phase of alcohol intake. The three clusters represent the brain activity of people with hazardous alcohol use, dependence symptoms, and lastly severe alcohol use.

For our overall clustering results in Fig. 4.36, the BIRCH algorithm was implemented. The results showed a maximum of 3 clusters that represent hazardous alcohol use, dependence symptoms, and severe alcohol use as a result of increasing alcohol consumption. The total number of scores recorded by the 17 participants in the three AUDIT domains are presented in the table 4.5. In this table, the last domain was modified from ”Harmful Alcohol use” to ”Severe alcohol use”. This is because the second consumption level in the AUDIT, ”hazardous or harmful risk consumption,” uses ”hazardous and harmful” terminology as a similar item. Based on the responses of the AUDIT, the maximum number of responses recorded is under the domain “hazardous alcohol use”, followed by “severe alcohol use”, and “dependence symptoms”. The highest score was reported for typical quantity, followed by frequency of heavy drinking, and lastly, frequency of drinking. From the clustering results in Fig. 4.36, It can be said that the red cluster represents this set of brain activity. The majority of participants have 5 or more drinks containing alcohol on a typical day when they are drinking, which can be classified as binge drinking. The frequency of drinking establishes the typical quantity of drinks on one occasion. In the AUDIT, this can be assessed as a period of either daily, weekly, or monthly. In this domain, individuals may project mild symptoms of AUD because their brain activity shows a higher degree of signal entanglement due to increasing alcohol consumption. The

data points are not confined as those in the blue and green cluster,s indicating a higher performance and motor coordination impairment compared to the people with dependence symptoms and severe alcohol use. In the domain of “dependence symptoms”, it can be observed that the need for a morning drink had the most number of responses, followed by impaired control, and increased salience over drinking. The green cluster in Fig. 4.36 represents this set of brain activity. This is to say that the majority of participants would normally need a first drink in the morning after a heavy drinking session. Impaired control over drinking normally refers to a person’s inability to abstain from drinking even if there is a possibility to do so. In this state, the subject could subconsciously choose to stop or continue drinking alcohol. It can be said that the individual could be having an imbalance between the excitatory and inhibitory neurons [95]. Increased salience of drinking means a person’s inability to perform what would be expected of them due to drinking. That is to say, an individual could prioritize consuming alcohol over other important activities. The results in 4.9 indicate that they normally stop rather than continue drinking and potentially start again the next morning. This is because increased salience of drinking had the lowest score compared to the other two contents. Although the subjects have varying responses to the effects of alcohol, as previously shown in the PSD analysis, it can be said that the majority of them project dependence symptoms caused by increasing the amount of alcohol consumption but are not entirely suffering from AUD. This is because the results obtained from the power spectrum were higher in the theta band compared to the alpha and beta bands. In this domain, individuals may project moderate symptoms of AUD which results in lesser reaction time and information processing due to higher doses of alcohol. Lastly, the majority of participants in the ‘severe alcohol use” domain reported “guilt after drinking” and “blackouts” equally, more than “concern from others”, and having “alcohol-related injuries”. The blue cluster Fig. 4.36 represents this set of brain activity. It can be said that alcoholics would normally prefer a blackout or have regrets about drinking heavily rather than having a morning drink. It is believed that the two aforementioned are directly influenced by impaired control over drinking. This is because if an individual chooses to continue drinking alcohol, they may end up ”blacked out” and have ”guilt after drinking” the next day. In this domain, individuals may project severe symptoms of AUD and are likely to have alcohol-related injuries if they engage more in alcohol consumption episodes. This may cause a need for an intervention from mental health professionals.

5. Conclusion and Future Directions

In conclusion, this research study has presented a model that correlates clinical assessments of SUD and brain-computer interface experiments with alcohol stimulus. Firstly, the extent of SUD was examined using data from the National Survey on Drug Use and Health online database that is based on DSM-5 criteria for SUD. Unsupervised machine learning was applied using three algorithms, k-means, hierarchical clustering, and DBSCAN. These algorithms result in consistent clustering of two major clusters, namely, respondents with SUD and those without. Based on the data visualization, the results are analyzed and interpreted from the point of view of machine learning. The interpretation based on numerical results is validated by a mental health professional. The study proceeds to analyze the severity of AUD from a group of seventeen male participants based on EEG and alcohol-drinking experiments using PSD and the BIRCH clustering algorithm. The data is recorded as they drink alcohol in increasing amounts. In this work, three types of results are analyzed that are, EEG spectral power, transitions from the initial to the final phase of alcohol intake, and lastly the overall brain activity of all the participants. The findings of the PSD indicate that more alcohol consumption results in more spectral power. A person with a higher frequency of alcohol consumption is less affected by its effects compared to those with lower consumption for the same amount of alcohol consumed. The transition-based clustering results are consistent with the PSD results because the interpretations indicate that people with a higher frequency of consumption have lesser motor control impairment and, therefore, are more coherent in their concentration level than those with a lesser frequency of consumption. For the overall clustering results, individuals with harmful alcohol use project mild symptoms of AUD, those with dependence symptoms project moderate symptoms of AUD, and lastly those with severe alcohol use project severe symptoms of AUD. The interpretation of these results corresponds with those from the PSD and the transition-based clustering. This approach contributes to mental health disorders through a computational model that automatically classifies people with or without SUD, and that is validated by a mental health professional. This tool can help supplement evaluations and assessments by mental health professionals, thus speeding up the diagnostics process, performing the necessary interventions accordingly, and how to manage interventions. In addition, an analysis of the severity of SUD for each of alcohol, cannabis, and cocaine is provided. The severity of AUD was analyzed through alcohol-drinking experiments because alcohol is a very chronic type of SUD. Lastly, this model is also able to discover patterns of other mental health disorders that may be related to the degree of severity of SUD or poly-SUD.

For future work, the results of this study will be improved by comparing multiple spectral fea-

tures to further assess the severity levels of AUD. Additionally, task-based EEG paradigms will be incorporated into future alcohol drinking experiments to allow for a more controlled assessment of cognitive and motor responses as alcohol intake increases. To complement EEG-based clustering, future studies will also include structured behavioral observations recorded in real-time during alcohol intake, such as visible changes in motor coordination, speech, or reaction times. These observations will be used to correlate the neural activity and enhance the interpretations. Furthermore, objective measures of alcohol levels, such as breathalyzer-based blood alcohol concentration (BAC) readings, will be integrated at regular intervals during the experiment. This will allow for more accurate tracking of intoxication levels across epochs and will strengthen the interpretation of clustering outcomes. Together, these additions will provide a more comprehensive understanding of the neurophysiological effects of alcohol use.

A. Feature Set – Alcohol Dataset

This appendix presents the features extracted from the NSDUH 2019 alcohol-related dataset. These variables were derived from participant survey responses and used in correlation and clustering analysis shown in Chapter 3.

ALCOHOL

These questions are about drinks of alcoholic beverages. Throughout these questions, by a "drink" we mean a can or bottle of beer, a glass of wine or a wine cooler, a shot of liquor, or a mixed drink with liquor in it. We are not asking about times when you only had a sip or two from a drink. Have you ever, even once, had a drink of any type of alcoholic beverage? Please do not include times when you only had a sip or two from a drink.

1. ALCEVER - EVER HAD DRINK OF ALCOHOLIC BEVERAGE
2. ALCTRY - AGE WHEN FIRST DRANK ALCOHOLIC BEVERAGE
3. ALCYFU - YEAR OF FIRST ALCOHOL USE - RECODE
4. ALCMFU - MONTH OF FIRST ALCOHOL USE - RECODE
5. ALCREC - TIME SINCE LAST DRANK ALCOHOLIC BEVERAGE
6. ALCYRTOT - TOTAL # OF DAYS USED ALCOHOL IN PAST 12 MONTHS
7. ALDAYPYR - NUMBER DAYS DRANK ALCOHOL IN PAST 12 MOS
8. ALDAYPMO - # DAYS PER MONTH DRANK ALCOHOL IN PAST 12 MOS
9. ALDAYPWK - # DAYS PER WEEK DRANK ALCOHOL IN PAST 12 MOS
10. ALCDAYS - # DAYS HAD ONE OR MORE DRINKS PAST 30 DAYS
11. AL30EST - BEST ESTIMATE # DAYS HAD ONE OR MORE DRINKS
12. ALCUS30D - USUAL # OF DRINKS PER DAY PAST 30 DAYS
13. ALCBNG30D - # DAYS HAD FOUR/FIVE OR MORE DRINKS PAST 30 DYS
14. RSKBNGDLY - RISK HAVING 4 OR 5 DRINKS NEARLY EVERY DAY
15. RSKBNGWK - RISK HAVING 5 OR MORE DRINKS ONCE OR TWICE A WEEK
16. RSKYFQDGR - GET A REAL KICK OUT OF DOING DANGEROUS THINGS
17. RSKYFQTES - LIKE TO TEST YOURSELF BY DOING RISKY THINGS
18. GRSKBNGDLY - RC-GREAT RISK: HAVE 4-5 ALC DRINKS PER DAY
19. GRSKBNGWK - RC-GREAT RISK: HAVE 5+ ALC DRINKS 1-2 TIMES PER WEEK

20. ALCLOTTM - SPENT MONTH/MORE GETTING/DRNKG ALC PAST 12 MOS

21. ALCGTOVR - MONTH+ SPENT GETTING OVER ALC EFFECTS PST 12 MOS

22. ALCLIMIT - SET LIMITS ON ALCOHOL USE PAST 12 MONTHS

23. ALCKPLMT - ABLE TO KEEP LIMITS OR DRANK MORE PAST 12 MOS

24. ALCNDMOR - NEEDED MORE ALC TO GET SAME EFFECT PST 12 MOS

25. ALCLSEFX - DRNKG SAME AMT ALC HAD LESS EFFECT PAST 12 MOS

26. ALCCUTDN - WANT/TRY TO CUT DOWN/STOP DRNKG PAST 12 MOS

27. ALCCUTEV - ABLE TO CUT/STOP DRNKG EVERY TIME PAST 12 MOS

28. ALCCUT1X - CUT DOWN OR STOP DRNKG AT LEAST ONCE PAST 12 MOS

29. ALCWD2SX - HAD 2+ ALC WITHDRAWAL SYMPTOMS PST 12 MOS

30. ALCWDSMT - HAD 2+ ALC WDRAW SYM AT SAME TIME PST 12 MOS

31. ALCEMOPB - ALC CAUSE PRBS WITH EMOT/NERVES PAST 12 MOS

32. ALCEMCTD - CONTD TO DRINK ALC DESPITE EMOT PRBS

33. ALCPHLPB - ANY PHYS PRBS CAUSED/WORSND BY ALC PST 12 MOS

34. ALCPHCTD - CONTD TO DRINK ALC DESPITE PHYS PRBS

35. ALCLSACT - LESS ACTIVITIES B/C OF ALC USE PAST 12 MOS

36. ALCSERPB - ALC CAUSE SERS PRBS AT HOME/WORK/SCH PST 12 MOS

37. ALCPDANG - DRNK ALC AND DO DANGEROUS ACTIVITIES PST 12 MOS

38. ALCLAWTR - DRNK ALC CAUSE PRBS WITH LAW PAST 12 MOS

39. ALCFMFPB - DRNK ALC CAUSE PRBS W/FAMILY/FRIENDS PST 12 MOS

40. ALCFMCTD - CONTD TO DRINK ALC DESPITE PRBS W/ FAM/FRNDS

B. Feature Set – Marijuana Dataset

This appendix presents the features used for analysis of the marijuana-related dataset.

MARIJUANA

The next questions are about marijuana and hashish. Marijuana is also called pot or grass. Marijuana is usually smoked, either in cigarettes, called joints, or in a pipe. It is sometimes cooked in food. Hashish is a form of marijuana that is also called "hash." It is usually smoked in a pipe. Another form of hashish is hash oil.

Have you ever, even once, used marijuana or hashish?

1. MJEVER - EVER USED MARIJUANA/HASHISH
2. MJAGE - AGE WHEN FIRST USED MARIJUANA/HASHISH
3. MJYFU - YEAR OF FIRST MARIJUANA/HASHISH USE - RECODE
4. MJMFU - MONTH OF FIRST MARIJUANA/HASHISH USE - RECODE
5. MJREC - TIME SINCE LAST USED MARIJUANA/HASHISH
6. MJYRTOT - TOTAL # OF DAYS USED MARIJUANA IN PAST 12 MOS
7. MRDAYPYR - # DAYS USED MARIJUANA/HASHISH PAST 12 MONTHS
8. MRDAYPMO - # DAYS PER MO USED MARIJUANA/HASHISH PAST 12 MOS
9. MRDAYPWK - # DAYS PER WEEK USED MARIJUANA/HASHISH PAST 12 MOS
10. MJDAY30A - # DAYS USED MARIJUANA/HASHISH PAST 30 DAYS
11. MR30EST - BEST EST. # DAYS USED MARIJUANA PAST 30 DAYS
12. RSKYFQDGR - GET A REAL KICK OUT OF DOING DANGEROUS THINGS
13. RSKYFQTES - LIKE TO TEST YOURSELF BY DOING RISKY THINGS
14. RKFQPBLT - WEAR A SEATBELT WHEN RIDE FRONT PASS SEAT OF CAR
15. GRSKMRJMON - RC-GREAT RISK: USE MARIJUANA ONCE A MONTH
16. GRSKMRJWK - RC-GREAT RISK: USE MARIJUANA 1-2 TIMES PER WEEK
17. DIFOBTMRJ - RC-MARIJUANA FAIRLY OR VERY EASY TO OBTAIN
18. MRJLOTTM - SPENT MONTH/MORE GETTING/USING MJ PAST 12 MOS
19. MRJGTOVR - MONTH+ SPENT GETTING OVER MJ EFFECTS PST 12 MOS

20. MRJLIMIT - SET LIMITS ON MARIJUANA USE PAST 12 MONTHS
21. MRJKPLMT - ABLE TO KEEP LIMITS OR USE MORE MJ PAST 12 MOS
22. MRJNDMOR - NEEDED MORE MJ TO GET SAME EFFECT PST 12 MOS
23. MRJLSEFX - USING SAME AMT MJ HAD LESS EFFECT PAST 12 MOS
24. MRJCUTDN - WANT/TRY TO CUT DOWN/STOP USING MJ PST 12 MOS
25. MRJCUTEV - ABLE TO CUT/STOP USING MJ EVERY TIME PST 12 MOS
26. MRJEMOPB - MJ CAUSE PRBS WITH EMOT/NERVES PAST 12 MOS
27. MRJEMCTD - CONTD USING MARIJUANA DESPITE EMOT PRBS
28. MRJPHLPB - ANY PHYS PRBS CAUSED/WORSND BY MJ PST 12 MOS
29. MRJPHCTD - CONTD TO USE MARIJUANA DESPITE PHYS PRBS
30. MRJLSACT - LESS ACTIVITIES B/C OF MJ USE PAST 12 MOS
31. MRJSERPB - MJ CAUSE SERS PRBS AT HOME/WORK/SCH PST 12 MOS
32. MRJPDANG - USING MJ AND DO DANGEROUS ACTIVITIES PST 12 MOS
33. MRJLAWTR - USING MJ CAUSE PRBS WITH LAW PAST 12 MOS
34. MRJFMFPB - USING MJ CAUSE PRBS W/FAMILY/FRIENDS PST 12 MOS
35. MRJFMCTD - CONTD TO USE MJ DESPITE PRBS W/ FAM/FRNDS

C. Feature Set – Cocaine Dataset

This appendix presents the features used for analysis of the cocaine-related dataset.

COCAINE

These questions are about cocaine, including all the different forms of cocaine such as powder, "crack," free base, and coca paste. Have you ever, even once, used any form of cocaine?

1. COCEVE - EVER USED COCAINE
2. COCAGE - AGE WHEN FIRST USED COCAINE
3. COCYFU - YEAR OF FIRST COCAINE USE - RECODE
4. COCMFU - MONTH OF FIRST COCAINE USE - RECODE
5. COCREC - TIME SINCE LAST USED COCAINE
6. COCYRTOT - TOTAL # OF DAYS USED COCAINE IN PAST 12 MONTHS
7. CCDAYPYR - # DAYS USED COCAINE PAST 12 MONTHS
8. CCDAYPMO - # DAYS PER MONTH USED COCAINE PAST 12 MONTHS
9. CCDAYPWK - # DAYS PER WEEK USED COCAINE PAST 12 MONTHS
10. COCUS30A - # DAYS USED COCAINE PAST 30 DAYS
11. CC30EST - BEST ESTIMATE # DAYS USED COCAINE PAST 30 DAYS
12. RSKCOCMON - RISK USING COCAINE ONCE A MONTH
13. RSKCOCWK - RISK USING COCAINE ONCE OR TWICE A WEEK
14. DIFGETCOC - HOW DIFFICULT TO GET COCAINE
15. RSKYFQDGR - GET A REAL KICK OUT OF DOING DANGEROUS THINGS
16. RSKYFQTES - LIKE TO TEST YOURSELF BY DOING RISKY THINGS
17. GRSKCOCMON - RC-GREAT RISK: USE COCAINE ONCE A MONTH
18. GRSKCOCWK - RC-GREAT RISK: USE COCAINE 1-2 TIMES PER WEEK
19. COCLOTTM - SPENT MONTH/MORE GETTING/USING COC PAST 12 MOS
20. COCGTOVR - MONTH+ SPENT GETTING OVER COC EFFECTS PST 12 MOS
21. COCLIMIT - SET LIMITS ON COCAINE USE PAST 12 MONTHS
22. COCKPLMT - ABLE TO KEEP LIMITS OR USED MORE COC PST 12 MOS

23. COCNDMOR - NEEDED MORE COC TO GET SAME EFFECT PST 12 MOS

24. COCLSEFX - USING SAME AMT COC HAD LESS EFFECT PAST 12 MOS

25. COCCUTDN - WANT/TRY TO CUT DOWN/STOP USING COC PAST 12 MOS

26. COCCUTEV - ABLE TO CUT/STOP USING COC EVERY TIME PAST 12 MOS

27. COCCUT1X - CUT DOWN/STOP USING COC AT LEAST 1X PST 12 MOS

28. COCFLBLU - WHEN CUT DOWN ON COC, FELT BLUE PAST 12 MONTHS

29. COCWD2SX - HAD 2+ COC WITHDRAWAL SYMPTOMS PST 12 MOS

30. COCWDSMT - HAD 2+ COC WDRAW SYM AT SAME TIME PAST 12 MOS

31. COCEMOPB - COC CAUSE PRBS WITH EMOT/NERVES PAST 12 MOS

32. COCEMCTD - CONTD TO USE COC DESPITE EMOT PRBS

33. COCPHLPB - ANY PHYS PRBS CAUSED/WORSND BY COC PST 12 MOS

34. COCPHCTD - CONTD TO USE COC DESPITE PHYS PRBS

35. COCLSACT - LESS ACTIVITIES B/C OF COC USE PAST 12 MOS

36. COCSERPB - COC CAUSE SERS PRBS AT HOME/WORK/SCH PST 12 MOS

37. COCPDANG - USING COC AND DO DANGEROUS ACTIVITIES PST 12 MOS

38. COCLAWTR - USING COC CAUSE PRBS WITH LAW PAST 12 MOS

39. COCFMFPB - USING COC CAUSE PRBS W/FAMILY/FRIENDS PST 12 MOS

40. COCFMCTD - CONTD TO USE COC DESPITE PRBS W/ FAM/FRNDS

D. Power spectral density graphs for all the 17 participants

The PSD results for the 17 subjects are shown in this chapter. To keep this document concise, these files are included in a zipped folder archive because they are too many. They can be downloaded from the attached compressed file using the following link: PSD Results for All 17 Participants

E. Approval Letter from Ministry of Health and Wellness

Approval letters from the Ministry of Health and Wellness (MOHW) under the Health Research and Development Division (HRDD). We attach the Initial and Final research permits.

PRIVATE BAG 0038
GABORONE
BOTSWANA
REFERENCE:

REPUBLIC OF BOTSWANA
MINISTRY OF HEALTH AND WELLNESS

TEL: (+267) 363 2500
FAX: (+267) 391 0647
TELEGRAMS: RABONGAKA
TELEX: 2818 CARE BD

REFERENCE NO: HPDME 13/18/1

25th August 2020

Health Research and Development Division

Notification of IRB Review: **New application**

Kaloso M. Tlotleng
P/bag 16
Palapye

Dear Kaloso M. Tlotleng
Lucky Mohutsiwa

Protocol Title: **MACHINE LEARNING MODEL TO CORRELATE BRAIN SIGNAL AND HUMAN BEHAVIOUR AFFECTED BY ADDICTION VERSION 1.0**

HRDD Approval Date:	25 August 2020
HRDD Expiration Date:	24 August 2021
HRDD Review Type:	Expedited Review
HRDD Review Determination:	Approved
Risk Determination:	Minimal risk

Thank you for submitting new application for the above referenced protocol. The permission is granted for Kaloso M. Tlotleng and Lucky Mohutsiwa to conduct the study.

This permit does not however give you authority to collect data from the selected sites without prior approval from the management. Consent from the identified individuals should be obtained at all times.

The research should be conducted as outlined in the approved proposal. Any changes to the approved proposal must be submitted to the Health Research and Development Division in the Ministry of Health for consideration and approval.

Furthermore, you are requested to submit at least one hardcopy and an electronic copy of the report to the Health Research, Ministry of Health and Wellness within 3 months of completion of the study. Approval is for academic fulfillment only. Copies should also be submitted to all other relevant authorities.

Continuing Review
In order to continue work on this study (including data analysis) beyond the expiry date, submit a Continuing Review Form for Approval at least three (3) months prior to the

protocol's expiration date. The Continuing Review Form can be obtained from the Health Research Division Office (HRDD), Office No. 7A.7 or Ministry of Health website: www.moh.gov.bw or can be requested via e-mail from Mr. Kgomotso Motlhanka, e-mail address: kgmmotlhanka@gov.bw As a courtesy, the HRDD will send you a reminder email about eight (8) weeks before the lapse date, but failure to receive it does not affect your responsibility to submit a timely Continuing Report form

Amendments

During the approval period, if you propose any change to the protocol such as its funding source, recruiting materials, or consent documents, you must seek HRDC approval before implementing it. Please summarize the proposed change and the rationale for it in the amendment form available from the Health Research Division Office (HRDD), Office No. 7A 7 or Ministry of Health website: www.moh.gov.bw or can be requested via e- mail from Mr. Kgomotso Motlhanka, e-mail address: kgmotlhanka@gov.bw . In addition submit three copies of an updated version of your original protocol application showing all proposed changes in bold or "track changes".

Reporting

Other events which must be reported promptly in writing to the HRDC include:

- Suspension or termination of the protocol by you or the grantor
- Unexpected problems involving risk to subjects or others
- Adverse events, including unanticipated or anticipated but severe physical harm to subjects.

If you have any questions please do not hesitate to contact Mr. K. Motlhanka at kgmmotlhanka@gov.bw, Tel +267-3632751. Thank you for your cooperation and your commitment to the protection of human subjects in research.

Yours sincerely

Ms S. Mosweunyane

for /PERMANENT SECRETARY

MINISTRY OF HEALTH AND WELLNESS
RESEARCH DIVISION
25 AUG 2020
P/BAG 0038
GABORONE
REPUBLIC OF BOTSWANA

PRIVATE BAG 0038
GABORONE
BOTSWANA
REFERENCE:

REPUBLIC OF BOTSWANA
MINISTRY OF HEALTH AND WELLNESS

TEL: (+267) 363 2500
FAX: (+267) 391 0647
TELEGRAMS: RABONGAKA
TELEX: 2818 CARE BD

REFERENCE NO: HPDME 13/18/1

17 September 2021

Health Research Development Committee

Principal Investigator: Kaloso Tlotleng
Notification of IRB Review: **Continuing Review**

Protocol Title: **MACHINE LEARNING MODEL TO CORRELATE BRAIN SIGNAL AND HUMAN BEHAVIOUR AFFECTED BY ADDITION**

Review Type: HRDD/Expedited
Review Date: 17 September 2021
Approval Date: 17 September 2021
Effective Date: 17 September 2021
Expiration Date: 16 September 2022

This certifies that the continuing review request for the protocol above was reviewed under review procedures. Approval is valid for a period of 1 year.

__X__ Study is continuing.
____ Accrual complete with treatment intervention and /or participant interviews/ surveys continuing.
____ Subject interventions/data collection ended on (date):_____
____ Open for analysis only. Expected end date:
____ Competed (including all analysis). Dated completed:_____
____ Cooperative Review
____ Other, Please describe:
____ The study has not been activated.

If you have any questions please do not hesitate to contact Ms D. Babini at dbabini@gov.bw, Tel +267-3632754 and Mr. K. Motlhanka at kgmmotlhanka@gov.bw, Tel +267-3632751. Thank you for your cooperation and your commitment to the protection of human subjects in research.

Yours sincerely
PP
Dr P.Masokwane
for /PERMANENT SECRETARY

MINISTRY OF HEALTH AND WELLNESS
RESEARCH DIVISION
17 SEP 2021
P/BAG 0038
GABORONE
REPUBLIC OF BOTSWANA

F. Internal Approval Letter from the office of the Vice Chancellor

BOTSWANA INTERNATIONAL UNIVERSITY OF SCIENCE & TECHNOLOGY

FACULTY OF ENGINEERING & TECHNOLOGY

BOTSWANA INTERNATIONAL UNIVERSITY OF SCIENCE & TECHNOLOGY
09 OCT 2020
RECEIVED
VICE CHANCELLOR'S OFFICE

MEMORANDUM

APPROVED
SIGNATURE: Osatolo
DATE: 21/10/2020

TO: Vice Chancellor

FROM: Research Assistants - Mechanical, Energy, and Industrial Engineering Department

Kaloso M. Tlotleng

Lucky O. Mohutsiwa

13/10/2020
Executive Coordinator
Advise
Osatolo
VC

THRU: DVC, Research, Development and Innovation 8/10/2020

DVC, Academic Affairs Dennis Segner 8/10/2020

Acting Dean, Faculty of Engineering and Technology Supported

HOD, Mechanical, Energy, and Industrial Engineering 28/09/2020

Prof. Jamisola (Supervisor) 25/09/2020

DATE: September 24, 2020

RE: APPLICATION FOR APPROVAL TO CONDUCT HUMAN STUDY EXPERIMENT

The above subject matter refers.

We would like to request permission to conduct experiments to analyze brain signals of human volunteers. These experiments are designed to gain more understanding on human emotion and addiction as recorded by a non-invasive brain-computer interface. This is part of our study for the MEng in Mechatronics and Industrial Instrumentation. The proposed experiments include:

1. Emotional Analysis Using Multiwave EEG Headset
2. Machine Learning Model to Correlate Brain Signal and Human Behavior Affected by Addiction.

BOTSWANA INTERNATIONAL UNIVERSITY OF SCIENCE & TECHNOLOGY
RECEIVED
06 OCT 2020
DEPUTY VICE CHANCELLOR RESEARCH DEVELOPMENT & INNOVATION
PRIVATE BAG 16
PALAPYE

G. Ethical Clearance Certificate from Human Ethics Research Committee

BOTSWANA INTERNATIONAL UNIVERSITY OF SCIENCE & TECHNOLOGY

ETHICAL CLEARANCE CERTIFICATE

Certificate Reference Number:	**HREC-002**
Project title:	**Correlating Addiction Classifications from Clinical Assessments and Brain-Computer Interface Experiments**
Principal Investigator:	**Mr. Kaloso Tloteng**
Period	**26/01/2022-25/02/2023**

On behalf of the **Human Ethics Research Committee**, I hereby give ethical approval in respect of the undertakings contained in the above-mentioned project and research instrument(s). Should any other instruments be used, these require separate authorization. The Researcher may therefore commence with the research as from the date of this certificate, using the reference number indicated above.

The **Human Ethics Research Committee** retains the right to:

1. Withdraw or amend this Ethical Clearance Certificate if;
    i. Any unethical principal or practices are revealed or suspected.
    ii. Relevant information has been withheld or misrepresented.
    iii. Regulatory changes of whatsoever nature so require.
    iv. The conditions contained in the Certificate have not been adhered to.
2. Request access to any information or data at any time during the course or after completion of the project.

The **Human Ethics Research Committee** wishes you well in your research.



26/01/2022

Prof. Goitseone Malumbela

Chairman, *Human Ethics Research Committee* **Date**

H. Approval from the BIUST Clinic under the Division of Student Affairs and Campus Life

MEMORANDUM

TO: MEng student; Department of Mechanical, Energy and Industrial Engineering; Mr Tlotleng

FROM: Manager Health Services



Ms. K. L. Ogopotse

DATE: May 03rd, 2022

RE: NURSE TO ASSIST IN CONDUCTING HUMAN STUDY EXPERIMENTS

The subject refers.

As per the memo dated 25 April 2022, kindly note that your request has been acceded to. The nurse will be availed for 2 hours in the morning and 2 hours in the afternoon per day for 2 weeks.

This is to ensure that the nurse will not have to leave the clinic for a prolonged period. The nurse will be expected to continue with his daily duties and provide support to your study at the same time.

Thank you.

Cc: Dean Faculty of Engineering and Technology
HOD mechanical engineering and industrial engineering
Supervisor, Prof Rodrigo Jamisola

I. Approval from the Directorate on Physical Planning and Campus Facilities

BOTSWANA INTERNATIONAL UNIVERSITY OF SCIENCE & TECHNOLOGY

Directorate of Physical Planning & Campus Facilities

MEMORANDUM

TO: HOD, Mechanical, Energy & Industrial Engineering
Prof. Rodrigo Jamisola Jr

FROM: Director (Ag), PPCF

(M.Mashumba)

DATE: May 6th, 2022 **REF:** BIUST/PPCF/FM/9323

SUBJECT: REQUEST FOR TWO MALE SECURITY GUARDS TO ASSIST IN CONDUCTING HUMAN STUDY EXPERIMENTS

Reference is made to your request on the above;

We hereby confirm that we will assist with the two security guards required during the days of the experiments.

Thank you.

J. Approved Consent form for the Alcohol drinking experiments

BIUST SMaRT Lab Consent Form

CORRELATING ADDICTION CLASSIFICATIONS FROM CLINICAL ASSESSMENTS AND BRAIN-COMPUTER INTERFACE EXPERIMENTS

INTRODUCTION

This proposed study will be used to understand the neurocognitive behavior of alcohol users based on DSM-5 criteria for substance use disorders. This will be done through an interaction between the subject and the Brain-Computer Interface (BCI) which harnesses the brain signals of the user. The brain signals will be used to make a prediction machine learning model to better understand what happens in the neurocognition of the addicted individual. In this study, we built a machine-learning model based on clinical data from a mental health professional's assessment. The next phase is to record and analyze brain signals to correlate them with the output from clinical data.

PURPOSE OF THE STUDY

The main aim of the research experiment is to model clinical assessment data available on the online database, then try to correlate it with brain signals. This study will be supplementary to the psychiatric evaluation.

ELIGIBILITY CRITERIA

To be eligible, the participant should answer '**Yes**' to question 1 and '**No**' to questions 2 to 9.

1. Have you ever consumed 2 liters of 6% alcohol or more in two hours before? Yes/No
2. Do you have any alcohol allergies? Yes/No
3. Have you ever had a bad reaction to alcohol? Yes/No
4. Do you suffer from any liver disease? Yes/No
5. Do you suffer from diabetes? Yes/No
6. Do you have any history of diabetes? Yes/No
7. Do you suffer from kidney disease? Yes/No
8. Do you have any history of kidney disease? Yes/No
9. Have you ever been diagnosed with Alcohol dehydrogenase deficiency? Yes/No
10. Participants should be males aged between 18 and 60 years and required to declare that they have no history of mental illnesses.
11. They should also have short hair.
12. They should also be based in BIUST at the time of the experiment.

STUDY PROCEDURE

1. Each volunteer will be given a maximum of two liters of beer with an alcohol content of around **6%.**
2. Brain signals are recorded throughout the experiment using a non-invasive EEG head module.
3. The rate of consumption will be based on the willingness of the volunteer to take the alcohol. The timing and the estimated amount of alcohol intake will be recorded at each instance of intake.
4. The experiment will take a maximum of **two hours** for each participant.
5. A BIUST nurse is to be present throughout the conduct of the experiment to monitor the vital signs of the participants.

ALTERNATIVE PROCEDURES

The participants are expected to come to the experiments with dry, unoiled hair.

RISKS OR DISCOMFORTS

No side effects are expected from this experiment. All participants will only be given alcoholic beverages in increasing amounts.

SAFETY AND COMPLIANCE

The device to be used is an Emotiv Epoc Flex EEG head module. The device is totally non-invasive and is used by researchers worldwide. Its document manual stated the product to be Federal Communications Commission (FCC) compliant. The following information is stated in the manual:

EMOTIV products are intended for research applications and personal use only. Our products are not sold as Medical Devices as defined in EU directive 93/42/EEC. Our products are not designed or intended to be used for the diagnosis or treatment of disease.

FCC ID Number 2ADIH-FLEX01 and IC ID Number: 12769A-FLEX01.

EMOTIV has undertaken testing and confirms:

This device complies with the radio equipment directive (2014/53/EU).

BENEFITS
Participants will benefit from this by being part of cutting-edge research and working with eager young researchers who wish to contribute to the betterment of this world.

NEW INFORMATION
In case of any new developments or news, the participant will be informed before those developments are implemented.

COSTS TO SUBJECTS AND COMPENSATION
No costs are expected from the participants.

VOLUNTARY PARTICIPATION
Participation in the experiment is by volunteering, the subjects will not be forced to do anything they are uncomfortable with.

RIGHT TO WITHDRAW
Participants have the right to withdraw from the experiment if they feel overwhelmed or if they have a good reason to quit (like health, school, or family issues). Data gathered from participants who withdrew will not be considered in the study. In case of a withdrawal, participants will be requested to stay under the supervision of a nurse until they are ready to be taken home.

PRIVACY, ANONYMITY, AND CONFIDENTIALITY
The identity and information of participants will be kept confidential throughout the experiment and any publication will not reveal any information about the participants.

FUTURE USE OF INFORMATION
The data collected will be used for further research and the results may be published in a paper. The participant's personal information will not be published.

WHO TO CONTACT

Prof Rodrigo Jamisola: jamisolar@biust.ac.bw
Kaloso Tlotleng: kaloso.tlotleng@studentmail.biust.ac.bw
Mobile number: +267 72 461 941
office no. : +267 493 1649

STATEMENT OF CONSENT

I, the undersigned,

Full name: _____

Age: _____

Gender: _____

Occupation: _____

Marital status: _____

hereby declare, that I give consent to participate in this research experiment, and I am aware of the risks associated with this participation but I am willing to take such risks on my own accord.

Signature: _____

Place: _____

Date: _____

WITNESS STATEMENT

I, the undersigned,

Full name: _____

Gender: _____

Occupation: _____

Marital status: _____

hereby declare, that I am a witness to the titled research experiment mentioned above.
It is within my best knowledge that the volunteer understands the rules and regulations presented herein, and will adhere to them.
I agree to be present throughout the conduct of the experiment, and monitor the health status of the participants.

Signature: _____

Place: _____

Date: _____

K. The Alcohol Use Disorders Identification Test used in the experiments to test the frequency and rate of use of alcohol for the participants

The Alcohol Use Disorders Identification Test: Self-Report Version

PATIENT: Because alcohol use can affect your health and can interfere with certain medications and treatments, it is important that we ask some questions about your use of alcohol. Your answers will remain confidential so please be honest. Place an X in one box that best describes your answer to each question.

Questions	0	1	2	3	4	
1. How often do you have a drink containing alcohol?	Never	Monthly or less	2-4 times a month	2-3 times a week	4 or more times a week	
2. How many drinks containing alcohol do you have on a typical day when you are drinking?	1 or 2	3 or 4	5 or 6	7 to 9	10 or more	
3. How often do you have six or more drinks on one occasion?	Never	Less than monthly	Monthly	Weekly	Daily or almost daily	
4. How often during the last year have you found that you were not able to stop drinking once you had started?	Never	Less than monthly	Monthly	Weekly	Daily or almost daily	
5. How often during the last year have you failed to do what was normally expected of you because of drinking?	Never	Less than monthly	Monthly	Weekly	Daily or almost daily	
6. How often during the last year have you needed a first drink in the morning to get yourself going after a heavy drinking session?	Never	Less than monthly	Monthly	Weekly	Daily or almost daily	
7. How often during the last year have you had a feeling of guilt or remorse after drinking?	Never	Less than monthly	Monthly	Weekly	Daily or almost daily	
8. How often during the last year have you been unable to remember what happened the night before because of your drinking?	Never	Less than monthly	Monthly	Weekly	Daily or almost daily	
9. Have you or someone else been injured because of your drinking?	No		Yes, but not in the last year		Yes, during the last year	
10. Has a relative, friend, doctor, or other health care worker been concerned about your drinking or suggested you cut down?	No		Yes, but not in the last year		Yes, during the last year	
					Total	

Bibliography

[1] Trevor W Robbins and Luke Clark. Behavioral addictions. Current opinion in neurobiology, 30:66–72, 2015.

[2] Joseph D Deak and Emma C Johnson. Genetics of substance use disorders: a review. Psychological medicine, 51(13):2189–2200, 2021.

[3] Kwok Kei Mak, Kounseok Lee, and Cheolyong Park. Applications of machine learning in addiction studies: A systematic review. Psychiatry research, 275:53–60, 2019.

[4] T Wing Lo, Jerf WK Yeung, and Cherry HL Tam. Substance abuse and public health: A multilevel perspective and multiple responses, 2020.

[5] American Psychiatric Association et al. Diagnostic and statistical manual of mental disorders (DSM-5®). American Psychiatric Pub, 2013.

[6] Asociación Psiquiátrica Americana APA. Diagnostic and statistical manual of mental disorders. The American Psychiatric Association, 2013.

[7] Jalie A Tucker, Susan D Chandler, and Katie Witkiewitz. Epidemiology of recovery from alcohol use disorder. Alcohol research: current reviews, 40(3), 2020.

[8] Elizabeth M Burnette, Steven J Nieto, Erica N Grodin, Lindsay R Meredith, Brian Hurley, Karen Miotto, Artha J Gillis, and Lara A Ray. Novel agents for the pharmacological treatment of alcohol use disorder. Drugs, 82(3):251–274, 2022.

[9] Michele K Bohm, Yong Liu, Marissa B Esser, Jessica B Mesnick, Hua Lu, Yi Pan, and Kurt J Greenlund. Binge drinking among adults, by select characteristics and state—united states, 2018. American journal of transplantation, 21(12):4084–4091, 2021.

[10] Alcohol Research and Current Reviews Editorial Staff. Drinking patterns and their definitions. Alcohol research: current reviews, 39(1):17, 2018.

[11] Substance Abuse, Mental Health Services Administration, et al. 2015 national survey on drug use and health. 2016.

[12] Håkan Källmén, Tobias H Elgán, Peter Wennberg, and Anne H Berman. Concurrent validity of the alcohol use disorders identification test (audit) in relation to alcohol use disorder (aud)

severity levels according to the brief dsm-5 aud diagnostic assessment screener. Nordic Journal of Psychiatry, 73(7):397–400, 2019.

[13] Shui-Hua Wang, Khan Muhammad, Jin Hong, Arun Kumar Sangaiah, and Yu-Dong Zhang. Alcoholism identification via convolutional neural network based on parametric relu, dropout, and batch normalization. Neural Computing and Applications, 32:665–680, 2020.

[14] American Psychiatric Association. Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition, Text Revision (DSM-5-TR™). American Psychiatric Association Publishing, Washington, DC, 2022.

[15] Digambar Puri, Rohit Chudiwal, Jaswantsing Rajput, Sanjay Nalbalwar, Anil Nandgaonkar, and Abhay Wagh. Detection of alcoholism from eeg signals using spectral and tsallis entropy with svm. In 2021 International Conference on Communication information and Computing Technology (ICCICT), pages 1–5. IEEE, 2021.

[16] Arti Anuragi and Dilip Singh Sisodia. Alcohol use disorder detection using eeg signal features and flexible analytical wavelet transform. Biomedical Signal Processing and Control, 52:384–393, 2019.

[17] Wajid Mumtaz, Pham Lam Vuong, Likun Xia, Aamir Saeed Malik, and Rusdi Bin Abd Rashid. Automatic diagnosis of alcohol use disorder using eeg features. Knowledge-Based Systems, 105:48–59, 2016.

[18] Wajid Mumtaz, Pham Lam Vuong, Aamir Saeed Malik, and Rusdi Bin Abd Rashid. A review on eeg-based methods for screening and diagnosing alcohol use disorder. Cognitive neurodynamics, 12:141–156, 2018.

[19] Zitong Wan, Rui Yang, Mengjie Huang, Nianyin Zeng, and Xiaohui Liu. A review on transfer learning in eeg signal analysis. Neurocomputing, 421:1–14, 2021.

[20] Lucky Odirile Mohutsiwa and Rodrigo S Jamisola. Eeg-based human emotion classification using combined computational techniques for feature extraction and selection in six machine learning models. In 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), pages 1095–1102. IEEE, 2021.

[21] Swati Aggarwal and Nupur Chugh. Review of machine learning techniques for eeg based brain computer interface. Archives of Computational Methods in Engineering, pages 1–20, 2022.

[22] M Srividya, S Mohanavalli, and N Bhalaji. Behavioral modeling for mental health using machine learning algorithms. Journal of medical systems, 42(5):1–12, 2018.

[23] Adane Nega Tarekegn, Krzysztof Michalak, and Mario Giacobini. Cross-validation approach to evaluate clustering algorithms: An experimental study using multi-label datasets. SN Computer Science, 1(5):1–9, 2020.

[24] Elizabeth Stevens, Dennis R Dixon, Marlena N Novack, Doreen Granpeesheh, Tristram Smith, and Erik Linstead. Identification and analysis of behavioral phenotypes in autism spectrum disorder via unsupervised machine learning. International journal of medical informatics, 129:29–36, 2019.

[25] Yankang Jing, Ziheng Hu, Peihao Fan, Ying Xue, Lirong Wang, Ralph E Tarter, Levent Kirisci, Junmei Wang, Michael Vanyukov, and Xiang-Qun Xie. Analysis of substance use and its outcomes by machine learning i. childhood evaluation of liability to substance use disorder. Drug and alcohol dependence, 206:107605, 2020.

[26] Zachary L Mannes, Dvora Shmulewitz, Ofir Livne, Malki Stohl, and Deborah S Hasin. Correlates of mild, moderate, and severe alcohol use disorder among adults with problem substance use: Validity implications for dsm-5. Alcoholism: Clinical and Experimental Research, 45(10):2118–2129, 2021.

[27] Carrie M Mintz, Sarah M Hartz, Sherri L Fisher, Alex T Ramsey, Elvin H Geng, Richard A Grucza, and Laura J Bierut. A cascade of care for alcohol use disorder: Using 2015–2019 national survey on drug use and health data to identify gaps in past 12-month care. Alcoholism: Clinical and Experimental Research, 45(6):1276–1286, 2021.

[28] Amie C Hayley, Con Stough, and Luke A Downey. Dsm-5 cannabis use disorder, substance use and dsm-5 specific substance-use disorders: Evaluating comorbidity in a population-based sample. European Neuropsychopharmacology, 27(8):732–743, 2017.

[29] Mustafa Al Sukar, Azzam Sleit, Abdullatif Abu-Dalhoum, and Bassam Al-Kasasbeh. Identifying a drug addict person using artificial neural networks. International Journal of Computer and Information Engineering, 10(3):611–616, 2016.

[30] Mohammad Kharabsheh, Omar Meqdadi, Mohammad Alabed, Sreenivas Veeranki, Ahmad Abbadi, and Sukaina Alzyoud. A machine learning approach for predicting nicotine dependence. International Journal of Advanced Computer Science and Applications, 10(3):179–184, 2019.

[31] Ali Ebrahimi, Uffe Kock Wiil, Kjeld Andersen, Marjan Mansourvar, and Anette Søgaard Nielsen. A predictive machine learning model to determine alcohol use disorder. In 2020 IEEE Symposium on Computers and Communications (ISCC), pages 1–7. IEEE, 2020.

[32] Woo-Young Ahn and Jasmin Vassileva. Machine-learning identifies substance-specific behavioral markers for opiate and stimulant dependence. Drug and alcohol dependence, 161:247–257, 2016.

[33] Substance Abuse, Mental Health Services Administration, et al. National survey on drug use and health. 2019.

[34] Randy Stinchfield, John McCready, Nigel E Turner, Susana Jimenez-Murcia, Nancy M Petry, Jon Grant, John Welte, Heather Chapman, and Ken C Winters. Reliability, validity, and classification accuracy of the dsm-5 diagnostic criteria for gambling disorder and comparison to dsm-iv. Journal of Gambling Studies, 32(3):905–922, 2016.

[35] Tulshi D Saha, S Patricia Chou, and Bridget F Grant. The performance of dsm-5 alcohol use disorder and quantity-frequency of alcohol consumption criteria: An item response theory analysis. Drug and Alcohol Dependence, 216:108299, 2020.

[36] KL Son, JS Choi, J Lee, SM Park, JA Lim, JY Lee, SN Kim, S Oh, DJ Kim, and JS Kwon. Neurophysiological features of internet gaming disorder and alcohol use disorder: a resting-state eeg study. Translational psychiatry, 5(9):e628–e628, 2015.

[37] Wajid Mumtaz, Nidal Kamel, Syed Saad Azhar Ali, Aamir Saeed Malik, et al. An eeg-based functional connectivity measure for automatic detection of alcohol use disorder. Artificial intelligence in medicine, 84:79–89, 2018.

[38] Jardel das C Rodrigues, Pedro P Rebouças Filho, Eugenio Peixoto Jr, Arun Kumar, and Victor Hugo C de Albuquerque. Classification of eeg signals to detect alcoholism using machine learning techniques. Pattern Recognition Letters, 125:140–149, 2019.

[39] Hamid Mukhtar, Saeed Mian Qaisar, and Atef Zaguia. Deep convolutional neural network regularization for alcoholism detection using eeg signals. Sensors, 21(16):5456, 2021.

[40] Leila Farsi, Siuly Siuly, Enamul Kabir, and Hua Wang. Classification of alcoholic eeg signals using a deep learning method. IEEE Sensors Journal, 21(3):3552–3560, 2020.

[41] Sachin Taran and Varun Bajaj. Rhythm-based identification of alcohol eeg signals. IET Science, Measurement & Technology, 12(3):343–349, 2018.

[42] Sunil Kumar Prabhakar and Harikumar Rajaguru. Alcoholic eeg signal classification with correlation dimension based distance metrics approach and modified adaboost classification. Heliyon, 6(12), 2020.

[43] Shelia Rahman, Tanusree Sharma, and Mufti Mahmud. Improving alcoholism diagnosis: comparing instance-based classifiers against neural networks for classifying eeg signal. In International Conference on Brain Informatics, pages 239–250. Springer, 2020.

[44] Deborah S Hasin, Charles P O'Brien, Marc Auriacombe, Guilherme Borges, Kathleen Bucholz, Alan Budney, Wilson M Compton, Thomas Crowley, Walter Ling, Nancy M Petry, et al. Dsm-5 criteria for substance use disorders: recommendations and rationale. American Journal of Psychiatry, 170(8):834–851, 2013.

[45] Substance Abuse et al. Key substance use and mental health indicators in the united states: Results from the 2019 national survey on drug use and health. 2020.

[46] Puneet Mishra, Alessandra Biancolillo, Jean Michel Roger, Federico Marini, and Douglas N Rutledge. New data preprocessing trends based on ensemble of multiple preprocessing techniques. TrAC Trends in Analytical Chemistry, 132:116045, 2020.

[47] Johura Khatun, Tamanna Azad, Samin Yeasar Seaum, Md Ismail Jabiullah, and Md Tarek Habib. Betel nut addiction detection using machine learning. In 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA), pages 1435–1440. IEEE, 2020.

[48] Raquel Iniesta, D Stahl, and Peter McGuffin. Machine learning, statistical learning and the future of biological research in psychiatry. Psychological medicine, 46(12):2455–2465, 2016.

[49] Amit K Chowdhry, Vinai Gondi, and Stephanie L Pugh. Missing data in clinical studies. International Journal of Radiation Oncology* Biology* Physics, 2021.

[50] Alma B Pedersen, Ellen M Mikkelsen, Deirdre Cronin-Fenton, Nickolaj R Kristensen, Tra My Pham, Lars Pedersen, and Irene Petersen. Missing data and multiple imputation in clinical epidemiological research. Clinical epidemiology, 9:157, 2017.

[51] Jie Cai, Jiawei Luo, Shulin Wang, and Sheng Yang. Feature selection in machine learning: A new perspective. Neurocomputing, 300:70–79, 2018.

[52] Jan Hauke and Tomasz Kossowski. Comparison of values of pearson's and spearman's correlation coefficient on the same sets of data. 2011.

[53] Peshawa Jamal Muhammad Ali, Rezhna Hassan Faraj, Erbil Koya, Peshawa J Muhammad Ali, and Rezhna H Faraj. Data normalization and standardization: a technical report. Mach Learn Tech Rep, 1(1):1–6, 2014.

[54] Ferath Kherif and Adeliya Latypova. Principal component analysis. In Machine Learning, pages 209–225. Elsevier, 2020.

[55] Faith Franzwa, Leia A Harper, and Kristen G Anderson. Examination of social smoking classifications using a machine learning approach. Addictive Behaviors, 126:107175, 2022.

[56] MA Syakur, BK Khotimah, EMS Rochman, and BD Satoto. Integration k-means clustering method and elbow method for identification of the best customer profile cluster. In IOP Conference Series: Materials Science and Engineering, volume 336, page 012017. IOP Publishing, 2018.

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

[58] Iqbal H Sarker. Machine learning: Algorithms, real-world applications and research directions. SN Computer Science, 2(3):1–21, 2021.

[59] Saptarshi Chakraborty, Debolina Paul, Swagatam Das, and Jason Xu. Entropy regularized power k-means clustering. arXiv preprint arXiv:2001.03452, 2020.

[60] Rena Nainggolan, Resianta Perangin-angin, Emma Simarmata, and Astuti Feriani Tarigan. Improved the performance of the k-means cluster using the sum of squared error (sse) optimized by using the elbow method. In Journal of Physics: Conference Series, volume 1361, page 012015. IOP Publishing, 2019.

[61] Deepshikha Aggarwal and Deepti Sharma. Application of clustering for student result analysis. International Journal of Recent Technology and Engineering (IJRTE) ISSN, pages 2277–3878, 2019.

[62] Chunhui Yuan and Haitao Yang. Research on k-value selection method of k-means clustering algorithm. J, 2(2):226–235, 2019.

[63] Mingrui Zhang. Use density-based spatial clustering of applications with noise (dbscan) algorithm to identify galaxy cluster members. In IOP Conference Series: Earth and Environmental Science, volume 252, page 042033. IOP Publishing, 2019.

[64] Derya Birant and Alp Kut. St-dbscan: An algorithm for clustering spatial–temporal data. Data & knowledge engineering, 60(1):208–221, 2007.

[65] Yanshan Wang, Yiqing Zhao, Terry M Therneau, Elizabeth J Atkinson, Ahmad P Tafti, Nan Zhang, Shreyasee Amin, Andrew H Limper, Sundeep Khosla, and Hongfang Liu. Unsupervised machine learning for the discovery of latent disease clusters and patient subgroups using electronic health records. Journal of biomedical informatics, 102:103364, 2020.

[66] Christian Lopez, Scott Tucker, Tarik Salameh, and Conrad Tucker. An unsupervised machine learning method for discovering patient clusters based on genetic signatures. Journal of biomedical informatics, 85:30–39, 2018.

[67] Xingyu Jiang, Jiayi Ma, Junjun Jiang, and Xiaojie Guo. Robust feature matching using spatial clustering with heavy outliers. IEEE Transactions on Image Processing, 29:736–746, 2019.

[68] Yanchi Liu, Zhongmou Li, Hui Xiong, Xuedong Gao, and Junjie Wu. Understanding of internal clustering validation measures. In 2010 IEEE international conference on data mining, pages 911–916. IEEE, 2010.

[69] SE Bleeker, HA Moll, EWet al Steyerberg, ART Donders, Gerarda Derksen-Lubsen, DE Grobbee, and KGM Moons. External validation is necessary in prediction research:: A clinical example. Journal of clinical epidemiology, 56(9):826–832, 2003.

[70] Tzu-Tsung Wong and Nai-Yu Yang. Dependency analysis of accuracy estimates in k-fold cross validation. IEEE Transactions on Knowledge and Data Engineering, 29(11):2417–2427, 2017.

[71] Tadayoshi Fushiki. Estimation of prediction error by using k-fold cross-validation. Statistics and Computing, 21(2):137–146, 2011.

[72] NJ Salkind. Root mean square error. Encyclopedia of Research Design, 2010.

[73] Tianfeng Chai and Roland R Draxler. Root mean square error (rmse) or mean absolute error (mae). Geoscientific Model Development Discussions, 7(1):1525–1534, 2014.

[74] Soraia M Alarcao and Manuel J Fonseca. Emotions recognition using eeg signals: A survey. IEEE transactions on affective computing, 10(3):374–393, 2017.

[75] Henry R Kranzler and Michael Soyka. Diagnosis and pharmacotherapy of alcohol use disorder: a review. Jama, 320(8):815–824, 2018.

[76] Sanne Verhoog, Jolien M Dopmeijer, Jannet M De Jonge, Claudia M Van Der Heijde, Peter Vonk, Rob HLM Bovens, Michiel R De Boer, Trynke Hoekstra, Anton E Kunst, Reinout W Wiers, et al. The use of the alcohol use disorders identification test–consumption as an indicator

of hazardous alcohol use among university students. European addiction research, 26(1):1–9, 2020.

[77] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, and Matti S. Hämäläinen. MEG and EEG data analysis with MNE-Python. Frontiers in Neuroscience, 7(267):1–13, 2013.

[78] Sabine Leske and Sarang S Dalal. Reducing power line noise in eeg and meg data via spectrum interpolation. Neuroimage, 189:763–776, 2019.

[79] Pranesh Krishnan, Sazali Yaacob, Annapoorni Pranesh Krishnan, Mohamed Rizon, and Chun Kit Ang. Eeg based drowsiness detection using relative band power and short-time fourier transform. J. Robotics Netw. Artif. Life, 7(3):147–151, 2020.

[80] Alaa Tharwat. Independent component analysis: An introduction. Applied Computing and Informatics, 17(2):222–249, 2021.

[81] Ahmad Al-Qerem, Faten Kharbat, Shadi Nashwan, Staish Ashraf, and Khairi Blaou. General model for best feature extraction of eeg using discrete wavelet transform wavelet family and differential evolution. International Journal of Distributed Sensor Networks, 16(3):1550147720911009, 2020.

[82] Asghar Zarei and Babak Mohammadzadeh Asl. Automatic seizure detection using orthogonal matching pursuit, discrete wavelet transform, and entropy based features of eeg signals. Computers in Biology and Medicine, 131:104250, 2021.

[83] Fahd A Alturki, Khalil AlSharabi, Akram M Abdurraqeeb, and Majid Aljalal. Eeg signal analysis for diagnosing neurological disorders using discrete wavelet transform and intelligent techniques. Sensors, 20(9):2505, 2020.

[84] Khalil AlSharabi, Yasser Bin Salamah, Akram M Abdurraqeeb, Majid Aljalal, and Fahd A Alturki. Eeg signal processing for alzheimer's disorders using discrete wavelet transform and machine learning approaches. IEEE Access, 10:89781–89797, 2022.

[85] Weikuan Jia, Meili Sun, Jian Lian, and Sujuan Hou. Feature dimensionality reduction: a review. Complex & Intelligent Systems, 8(3):2663–2693, 2022.

[86] Basna Mohammed Salih Hasan and Adnan Mohsin Abdulazeez. A review of principal component analysis algorithm for dimensionality reduction. Journal of Soft Computing and Data Mining, 2(1):20–30, 2021.

[87] Wajid Mumtaz, Pham Lam Vuong, Likun Xia, Aamir Saeed Malik, and Rusdi Bin Abd Rashid. An eeg-based machine learning method to screen alcohol use disorder. Cognitive neurodynamics, 11(2):161–171, 2017.

[88] Jack T Waddell. Age-varying time trends in cannabis-and alcohol-related risk perceptions 2002–2019. Addictive behaviors, 124:107091, 2022.

[89] Nathaniel A Dell, Sweta Prasad Srivastava, Michael G Vaughn, Christopher Salas-Wright, Audrey Hang Hai, and Zhengmin Qian. Binge drinking in early adulthood: A machine learning approach. Addictive Behaviors, 124:107122, 2022.

[90] Leah Wetherill, Arpana Agrawal, Manav Kapoor, Sarah Bertelsen, Laura J Bierut, Andrew Brooks, Danielle Dick, Michie Hesselbrock, Victor Hesselbrock, Daniel L Koller, et al. Association of substance dependence phenotypes in the coga sample. Addiction biology, 20(3):617–627, 2015.

[91] Charlotte Pscherer, Moritz Mückschel, Lena Summerer, Annet Bluschke, and Christian Beste. On the relevance of eeg resting theta activity for the neurophysiological dynamics underlying motor inhibitory control. Human Brain Mapping, 40(14):4253–4265, 2019.

[92] Brian Scally, Melanie Rose Burke, David Bunce, and Jean-Francois Delvenne. Resting-state eeg power and connectivity are associated with alpha peak frequency slowing in healthy aging. Neurobiology of aging, 71:149–155, 2018.

[93] Madhavi Rangaswamy and Bernice Porjesz. Understanding alcohol use disorders with neuroelectrophysiology. Handbook of clinical neurology, 125:383–414, 2014.

[94] Hernán Díaz, Fernando Maureira Cid, Jaime Otárola, Roberto Rojas, Oscar Alarcón, and Lucio Cañete. Eeg beta band frequency domain evaluation for assessing stress and anxiety in resting, eyes closed, basal conditions. Procedia Computer Science, 162:974–981, 2019.

[95] Eduardo López-Caneda, Fernando Cadaveira, Angeles Correas, Alberto Crego, Fernando Maestú, and Socorro Rodríguez Holguín. The brain of binge drinkers at rest: alterations in theta and beta oscillations in first-year college students with a binge drinking pattern. Frontiers in behavioral neuroscience, 11:168, 2017.

[96] Rosa Jurado-Barba, Ana Sion, Andrés Martínez-Maldonado, Isabel Domínguez-Centeno, Julio Prieto-Montalvo, Francisco Navarrete, María Salud García-Gutierrez, Jorge Manzanares, and Gabriel Rubio. Neuropsychophysiological measures of alcohol dependence: can we use eeg in the clinical assessment? Frontiers in Psychiatry, 11:523188, 2020.

[97] Nathan Didier, Ashley Vena, Abigayle R Feather, Jon E Grant, and Andrea C King. Holding your liquor: Comparison of alcohol-induced psychomotor impairment in drinkers with and without alcohol use disorder. Alcohol: Clinical and Experimental Research, 47(6):1156–1166, 2023.

[98] Natalia Almeida-Antunes, Alberto Crego, Carina Carbia, Sónia S Sousa, Rui Rodrigues, Adriana Sampaio, and Eduardo López-Caneda. Electroencephalographic signatures of the binge drinking pattern during adolescence and young adulthood: a prisma-driven systematic review. NeuroImage: Clinical, 29:102537, 2021.