EEG-BASED MENTAL STATE CLASSIFICATION: A RESPONSIBLE AI APPROACH

HAMSE ELMI

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
BACHELOR OF SCIENCE IN COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE

DEPARTMENT OF
COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

STUDENT NUMBER

2023232

COMMITTEE

dr. Silvy Collin
Lisa Lepps

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

June 24, 2024

WORD COUNT

6020

ACKNOWLEDGMENTS

I would like to express my heartfelt appreciation and thanks to everyone who has helped me over the years to reach this point. Your support has been invaluable, and I am truly grateful for all the assistance I've received. A special thank you goes to my supervisor, Silvy, for always being there and ready to provide guidance whenever I needed it.

EEG-BASED MENTAL STATE CLASSIFICATION: A RESPONSIBLE AI APPROACH

HAMSE ELMI

Abstract

The rapid advancement of machine learning technologies presents new opportunities for understanding and predicting human mental states through electroencephalography (EEG) data analysis. However, the ethical deployment of these technologies remains a significant challenge. This research investigates how machine learning algorithms can be utilized to predict mental states from EEG data while ensuring responsible AI practices, including transparency, fairness, safety, and privacy. The primary goal of this study is to bridge the gap between EEG-based mental state classification and responsible AI practices. Using the test-retest resting and cognitive state EEG dataset collected by Wang, Duan, Dong, Ding, and Lei (2022a), various machine learning models, including Random Forest, Support Vector Machines (SVM), Gradient Boosting Machines, and Neural Networks, were evaluated for their accuracy in predicting mental states. A distinguishing feature of our method is the integration of the SHERPA approach for electrode selection through SHAP for model interpretability. This ensures that the models are not only accurate but also transparent and privacy-preserving. Our findings indicate that the Random Forest model consistently outperforms other models, achieving the highest accuracy in predicting mental states. This study demonstrates the feasibility of developing ethically responsible AI systems for EEG-based mental state classification, contributing to advancements in cognitive science and offering practical applications in fields such as healthcare, education, and human-computer interaction. Future research should focus on expanding dataset diversity and enhancing model interpretability to further improve the ethical deployment of these technologies.

1 DATA SOURCE, ETHICS, CODE, AND TECHNOLOGY STATEMENT

Data Source: The test-retest resting and cognitive state EEG dataset Wang et al. (2022a) was acquired from OpenNEURO through an online request. The data is anonymised, although the raw data contains features such as age, gender, weight, and height, which I deleted from the processed data used for the analysis. The dataset owner has consented to its use for secondary data analysis, adhering to OpenNEURO's data use agreement. All figures included in this thesis were generated from the code I wrote for the analysis. I took inspiration from a YouTube series Anwer (2021) explaining how EEG-based data analysis works, after which I wrote the code independently and adapted it for my use case within this analysis. The thesis code can be accessed through the GitHub repository hamseelmi199 (2024). A generative language model OpenAI (2023) was employed to enhance the clarity of the content, focusing on paraphrasing, spell checking, and grammar. No typesetting tools or services beyond this language model were used.

2 INTRODUCTION

The rapid advancement of machine learning technologies unlocks new possibilities for understanding and predicting human mental states. One of the spearheading domains leading mental state classification is through electroencephalography (EEG) data analysis. However, it is crucial to approach this progress responsibly, especially because we are in an age where scepticism about the use of artificial intelligence is rife and data collection by both private companies and governmental organisations is being scrutinised Coin, Mulder, and Dubljević (2020). Our research aims to bridge the gap between the domains of EEG-based mental state classification and responsible AI practices by integrating ethical considerations, transparency, fairness, safety, and privacy into the design and evaluation of EEG-based classification models. By doing so, it seeks to contribute to the development of artificial intelligence technologies that are not only accurate and effective but also ethical, trustworthy, and beneficial to society Coin et al. (2020).

This research utilizes EEG-based signals from the test-retest resting and cognitive state EEG dataset collected by Wang et al. (2022a), aiming to explore the neural correlates of mental states through advanced data analysis techniques. This research aims to address the gap between EEG-based mental state classification and responsible AI practices. By integrating ethical considerations, transparency, fairness, safety, and privacy into the design and evaluation of EEG-based classification models, I seek to contribute to the development of artificial intelligence technologies that are not only accurate and effective but also ethical, trustworthy, and beneficial to society Coin et al. (2020). Through this interdisciplinary approach, my research aims to advance both the scientific understanding of EEG-based mental state classification and the practical application of responsible AI principles in real-world contexts.

By uncovering the neural correlates of specific mental states through EEG data analysis, this research aims to optimise processes reliant on understanding mental engagement. Potential applications span various domains, including enhancing user experiences in entertainment and healthcare, improving performance in training environments, and refining educational settings Apicella et al. (2022).

By contributing to the broader field of cognitive science and pinpointing neural correlates, this study seeks to advance the field by creating an accurate prediction model that adheres to responsible AI practices. This involves minimizing data collection to essential aspects for understanding mental states and implementing various responsible AI practices that focus on producing artificial intelligence that is explainable, fair, safe, and privacy-preserving (Molnar, 2023). Thus, by incorporating ethical considerations and responsible AI practices throughout the model development and evaluation process, this research aims to advance the field while maintaining ethical standards and respecting user privacy (Molnar, 2023).

As mentioned earlier, the aim of this research is to explore the potential of machine learning algorithms in predicting mental states using EEG data while incorporating and integrating responsible AI practices to ensure ethical considerations.

RESEARCH QUESTION

- How can machine learning algorithms be effectively utilized to predict mental states based on EEG data while ensuring responsible AI practices are upheld throughout the model development and evaluation process?

We will answer the above question by taking into account the following sub-questions which all in some form address the challenges of optimising model performance while adhering to ethical guidelines

Sub-questions

- What ethical considerations need to be taken into account when working with EEG data for mental state classification?
- How can model interpretability techniques such as SHAP contribute to understanding the decision-making process of the classification model?

- How do hyper-parameter optimization and cross-validation techniques impact the model's performance and generalizability in predicting mental states accurately?
- How can the classification model be fine-tuned to balance accuracy and ethical considerations, ensuring responsible AI practices are maintained throughout the project?

3 RELATED WORK

3.1 *EEG and Mental States*

Mental states refer to different levels of consciousness and cognitive functioning that a person experiences. These include emotional states like happiness, sadness, anger, and fear; cognitive states involving levels of attention, memory, decision-making, and problem-solving abilities; and consciousness covering states of being awake, asleep, or in altered states like meditation or anaesthesia Hu, Li, Sun, and Ratcliffe (2018). Understanding and predicting mental states with EEG involves analysing brain wave patterns to infer what someone might be feeling, thinking, or experiencing at any given time. For instance, specific EEG patterns are linked to different sleep stages, levels of awareness, and emotional responses Peng et al. (2019).]Electroencephalography (EEG) is a non-invasive method for recording the brain's electrical activity. It involves placing electrodes on the scalp to detect brain waves, the electrical impulses produced by neurons communicating. EEG is widely used in both clinical and research settings for diagnosing neurological conditions like epilepsy, studying brain functions, and monitoring brain activity in various mental states Suhail, Indiradevi, Suhara, Poovathinal, and Ayyappan (2022).

Mental states refer to different levels of consciousness and cognitive functioning that a person experiences. These include emotional states like happiness, sadness, anger, and fear; cognitive states involving levels of attention, memory, decision-making, and problem-solving abilities; and consciousness covering states of being awake, asleep, or in altered states like meditation or anaesthesia Hu et al. (2018). Understanding and predicting mental states with EEG involves analysing brain wave patterns to infer what someone might be feeling, thinking, or experiencing at any given time. For instance, specific EEG patterns are linked to different sleep stages, levels of awareness, and emotional responses Peng et al. (2019).

3.2 *Responsible AI*

Responsible AI encompasses practices and guidelines to ensure that artificial intelligence technologies are developed and deployed in ways that are ethical, fair, and beneficial to society. Key aspects of responsible AI include:

- **Explainability**: Making the decisions and outputs of artificial intelligence systems understandable. For artificial intelligence models analysing EEG data, this means providing clear, comprehensible insights into how they draw conclusions about mental states. This transparency is vital for building trust and ensuring users and stakeholders can grasp the reasoning behind artificial intelligence predictions (Molnar, 2023).

- **Fairness**: Ensuring the technology does not create or perpetuate biases or discrimination. When it comes to analysing EEG data, it is essential that artificial intelligence models do not favour certain groups over others based on attributes like age, gender, ethnicity, or neurological diversity. Achieving fairness involves using diverse and representative datasets and continuously monitoring and mitigating biases in the artificial intelligence models Ryali, Zhang, de los Angeles, Supekar, and Menon (2024).

- **Safety**: Ensuring the technology does not cause harm to users or society. For EEG-based artificial intelligence systems, this means rigorously testing the models to prevent erroneous predictions that could negatively impact mental health diagnoses or treatments. Safety measures include robust error handling, validation of model predictions, and ensuring the technology works reliably in different environments Sylvester, Sagehorn, Gruber, Atzmueller, and Schöne (2024).

- **Privacy**: The ability for a person to determine how their personal information is used. This is a critical concern, especially with sensitive data like EEG recordings, which can reveal intimate details about an individual's mental states and cognitive functions. There is growing awareness that brain-computer interface technology could breach privacy and safety regulations, particularly with brain-to-brain interfacing technology, which could allow information to be obtained without the knowledge of one or both parties Coin et al. (2020). Yue (2023) raises questions about whether individuals should have a right to brain privacy. To address these concerns, privacy-preserving practices must include strict data protection measures. These measures include anonymizing data, implementing strong security protocols, and ensuring compliance with data protection regulations such as the European Union's General Data Protection Regulation European Union (2016). Users should have control over their data, understand how it is being used, and have the ability to opt out if desired.

3.3 *Binary Classification Studies:*

The study of mental states using EEG data has gained significant attention due to its potential applications in various fields such as education, healthcare, and human-computer interaction. EEG signals offer valuable insights into brain activity and can be utilized to infer mental states like attention, working memory, and mental workload Suhail et al. (2022). Previous research has demonstrated the utility of machine learning models in accurately predicting these mental states based on EEG data. Several studies have focused on binary classification of mental states using EEG data. Here is an explanation of the methods used within these studies:

- Liu, Chiang, and Chu (2013):
  - Method: EEG power spectral density features and a Support Vector Machine (SVM) classification model.
  - Explanation: Power spectral density measures the power of different frequency components of the EEG signal. SVM is a supervised machine learning model that finds the hyperplane that best separates different classes of data.
  - Result: Achieved an average accuracy of 76.82% in attention recognition.
- Hu et al. (2018):
  - Method: Correlation-based feature selection and a k-Nearest Neighbour (KNN) classifier.
  - Explanation: Correlation-based feature selection identifies the most relevant features by evaluating the correlation between features and the target variable. KNN classifies data points based on the 'k' closest training examples in the feature space.
  - Result: Achieved a classification rate of 80.84% in detecting attention from EEG signals.
- Acı, Kaya, and Mishchenko (2019):
  - Method: EEG spectral features and SVM.

  - Explanation: Spectral features are derived from the frequency domain representation of the EEG signals. SVM, as previously mentioned, is used for classification.
  - Result: Reported a maximum accuracy of 96.70% in attention recognition.

- Peng et al. (2019):
  - Method: Hilbert-Huang Transform (HHT) and Spectral entropy with SVM.
  - Explanation: HHT is a method to analyze non-stationary signals like EEG. Spectral entropy measures the complexity or disorder within a signal. SVM is then used to classify these features.
  - Result: Achieved an average accuracy of 84.80% in discriminating the mental states of attentiveness and relaxation.

- Maghsoudi and Shalbaf (2021):
  - Method: Various brain connectivity features and SVM.
  - Explanation: Brain connectivity features represent the functional connections between different regions of the brain. SVM is used to classify these features.
  - Result: Achieved an accuracy of 89.2% in classifying mental states during arithmetic operations.

- Toa, Sim, and Tan (2021):
  - Method: Convolution Attention Memory Neural Network model.
  - Explanation: This deep learning model combines convolutional layers (for feature extraction), attention mechanisms (to focus on important features), and memory components (to retain important information) for classification tasks.
  - Result: Obtained 92.5% accuracy in classifying subjects' EEG signals during attentive or inattentive behaviours.

- Zhang et al. (2022):
  - Method: EEG power spectral features and a back propagation neural network classifier.
  - Explanation: Power spectral features are used to represent the frequency components of the EEG signals. A back propagation neural network is a type of neural network that adjusts its weights based on the error of the output, allowing it to learn complex patterns.

– Result: Achieved up to 91.1% accuracy in detecting alertness levels.

Each of these methods employs different techniques to extract meaningful features from EEG data and classify them into specific mental states, demonstrating various degrees of accuracy in their results. However, despite the high accuracies reached in these studies, something they lack is adherence to responsible AI practices. Due to the sensitive nature of EEG data, it is important that this data is dealt with responsibly Coin et al. (2020).

3.4 *Introduction to SHERPA*

SHERPA represents a significant advancement in EEG analysis, particularly in studying event-related potentials (ERPs), which are brain responses directly linked to specific sensory, cognitive, or motor events. To enhance transparency and reliability, SHERPA leverages explainable AI, providing human-understandable explanations for its findings. It employs the following explainable AI method: SHapley Additive exPlanations (SHAP), a method that identifies critical aspects of EEG data such as specific time ranges or electrodes.

Through this SHERPA offers an objective approach to analysing ERPs, enhancing sensitivity and specificity. This precision enables researchers to gain a deeper understanding of neural processes Sylvester et al. (2024). By focusing on key data features, SHERPA enables a more precise understanding of neural processes associated with sensory, cognitive, and motor events, supporting scientifically robust and ethically sound research practices in EEG analysis.

3.5 *Insights into Brain Function and Behaviour*

Recent studies have explored the intricate relationship between brain function and behaviour using EEG and artificial intelligence techniques. For instance, the silent recitation of song lyrics engages language processing areas in the brain, necessitating the activation of regions in the left hemisphere, particularly the temporal and frontal areas Apicella et al. (2022). Furthermore, research utilizing artificial intelligence and extensive fMRI datasets has uncovered gender differences in brain function, especially within the default mode network, striatum, and limbic network. These findings enhance our understanding of how sex influences brain function and behaviour Ryali et al. (2024).

Additionally, the effects of ageing on the brain have been well-documented. Ageing leads to brain shrinkage, vascular and cognitive changes, and increases the risks of stroke, white matter lesions, dementia, and memory issues Peters (2006). Collectively, these studies contribute significantly to our understanding of brain function and behaviour and underscore the diverse applications of EEG data analysis in neuroscience and cognitive science.

4 METHOD

4.1 *Dataset*

As mentioned earlier, this research utilised EEG-based signals from the test-retest resting and cognitive state EEG dataset collected by Wang et al. (2022a). This dataset was obtained from 60 participants in a controlled laboratory environment, where they wore a 61-channel EEG cap. The dataset includes EEG recordings from resting states (eyes open and eyes closed) and subject-driven cognitive tasks (memory, music, subtraction) at three different time points: the initial session, a session 90 minutes later, and a session one month later Wang, Duan, Dong, Ding, and Lei (2022b). In our research, we focused on the music cognitive task, using all the data from the three time points. In addition to EEG data, the dataset also includes behavioural data related to demographics, sleep, emotion, mental health, and self-generated thoughts (mind-wandering). This rich dataset allows for the investigation of intra- and inter-session variability in cognitive states and EEG changes at high temporal resolution. For our use case, we utilized the Mini NYC-Q behavioural test Gorgolewski et al. (2014), which assesses and provides insights into the participant's mental state during different cognitive states and activities, combined with the EEG data obtained from the participants. The Mini NYC-Q test posed several questions regarding the participants' mental state which can be seen in Table 1.

Although not taken into account within the analysis, the age range of the total number of participants was from 18 to 28, with 27 males and 33 females. The other physical attributes in the dataset, namely height and weight were not considered relevant for this study.

4.1.1 *Preprocessing*

Due to the test-retest nature of the dataset, we possess EEG-based data from the same participants, collected under identical circumstances across multiple time intervals. This longitudinal aspect is invaluable for training

Question 1:	I thought about something positive
Question 2:	I thought about something negative
Question 3:	My thoughts involved future events
Question 4:	My thoughts involved past events
Question 5:	My thoughts involved myself
Question 6:	My thoughts involved other people
Question 7:	My thoughts involved my surroundings
Question 8:	I was fully awake
Question 9:	My thoughts were in the form of images
Question 10:	My thoughts were in the form of words
Question 11:	My thoughts were more specific than vague
Question 12:	My thoughts were intrusive

Table 1: Questions Used in This Research From the Mini New York Cognition Questionnaire (Mini Nyc-Q).

prediction classifiers. The dataset encompasses both raw and preprocessed data, with the latter having gone through a series of preprocessing techniques by Wang et al. for optimal analysis 2022b, serving as the foundation for my research.

The already implemented preprocessing steps included:

- Re-referencing and Filtering: The raw EEG data was initially re-referenced to a common average and filtered using a symmetric Finite Impulse Response filter within EEGLAB Delorme and Makeig (2004), restricting the frequency band to 0.3–45 Hz. This further culminated in the processed data being stored in a proprietary set file that is to be used within EEGLAB.

- Channel Inspection and Handling Missing Data: Visual inspection of EEG signals was conducted to identify problematic channels, defined as those with a problematic trial ratio of $1/3$. Fortunately, no such channels were detected. Any missing data were addressed through linear interpolation, replacing them with the average of neighbouring channels.

- Epoch Segmentation and Bad Segment Screening: Segments of 4-second epochs were extracted from the EEG signal. Each epoch underwent manual screening to detect bad segments, though none were identified in this dataset.

- Artefact Identification with Independent Component Analysis (ICA): Independent component analysis (ICA) was employed within EEGLAB to identify artefacts such as eye blinks and movements during the eyes-open state. On average, 0.978 (±0.146) components per EEG session were flagged as artefacts.

The behavioural data collected was stored in a file containing all the results as values within a wider TSV (tab-separated values) file. However, some of the behavioural data needed for my analysis was missing. Specifically, 3 participants did not take part in the Mini NYC-Q at all. This resulted in only 57 participants being used in my research. The further preprocessing of this data implemented for my model is expanded upon below.

4.2 *Responsible AI practices*

Within our research, we embrace the principles of responsible AI, which necessitate careful consideration of measures relevant to various aspects within this framework.

4.2.1 *Explainability*

Adhering to the principles of explainability in responsible AI is crucial. We prioritize the model's interpretability, facilitated through an analysis that explains the significance of EEG electrodes in model accuracy. Using the SHAP (SHapley Additive exPlanations) method, we conduct a thorough global analysis of the model, identifying the roles of all the EEG electrodes in its predictive accuracy Molnar (2023).

4.2.2 *Fairness*

Upholding fairness requires a careful approach to dataset curation, ensuring that biases or discriminatory tendencies are avoided. This involves excluding extraneous variables such as age, gender, height, and weight from the processed dataset. While acknowledging the potential usefulness of gender and age segmentation in understanding brain activity nuances, their omission from this research underscores our commitment to equitable treatment and bias mitigation.

4.2.3 *Safety*

Ensuring safety within responsible AI involves rigorous testing of model parameters and robust error mitigation strategies. We validate model predictions thoroughly to prevent erroneous outcomes that could compromise mental health diagnoses or therapeutic interventions.

4.2.4 *Privacy*

Ensuring adherence to the privacy clause of responsible AI demands an approach that takes into account data minimization and purpose limitation. We selectively filter the initially collected EEG electrodes to encompass solely those specific to music-related tasks Scrivener and Reader (2022), thus confining data collection to its intended scope. Subsequent refinement, informed by insights from the SHAP explainability analysis, narrows down data collection to the top 20 electrodes. This process protects user privacy by avoiding the collection of unnecessary data that could be used for other purposes, hence preventing unjustified intrusions on user privacy.

4.3 *Development cycle*

As previously stated, we are making use of the preprocessed data and the behavioural data file. The model development cycle entails the following steps: filtering the behavioural data file, organizing the directory containing the data files, assigning the correct target labels to the data files. And the following steps in a loop for each Mini NYC-Q question: performing exploratory data analysis (EDA), data cleaning, feature selection, machine learning models classification, SHAP analysis, and machine learning models classification based on the most optimal SHAP values.

Recognizing the nuanced degrees of variability in mental states and EEG data, we will divide the Mini NYC-Q labels into two distinct classes: positive and negative. This classification will enrich our model's capacity for nuanced and accurate prediction. The classes will consist of halves, with "pos", positive, corresponding to answers to each question being between 10 and 5, and "neg", negative, corresponding to answers between 4 and 0.

4.3.1 *Data Filtering*

As mentioned earlier, the initial dataset contains a wealth of information, not all of which is important to the analysis. Therefore, the first step is to filter this dataset to extract only the necessary details, specifically the participant IDs and their responses to the Mini NYC-Q questions related to the music cognitive task.

The filtering process involves:

1. **Reading the Data:** The raw data is loaded from a CSV file.
2. **Identifying Relevant Columns:** Columns containing keywords "music" and "mini" in their names are identified, ensuring that only columns relevant to the music cognitive task are selected.

Question of Mini NYC-Q	Pre-Balancing	Post-Balancing
1	93P – 78N	84P – 84N
2	41P – 130N	130P – 130N
3	83P – 88N	88P – 88N
4	87P – 84N	87P – 87N
5	113P – 58N	113P – 113N
6	119P – 52N	119P – 119N
7	63P – 108N	108P – 108N
8	109P – 62N	109P – 109N
9	111P – 60N	111P – 111N
10	47P – 124N	124P – 124N
11	106P – 65N	106P – 106N
12	80P – 91N	91P – 91N

Table 2: Pre-Balancing and Post-Balancing Results for Mini NYC-Q. Note: P stands for Positive, and N stands for Negative.

3. **Creating a Filtered Dataset:** A new dataset is created containing only these relevant columns.

4. **Saving the Filtered Data:** This refined dataset is saved into a new CSV file, which serves as the basis for further processing.

This step ensures that only necessary data is retained, simplifying subsequent analysis steps and reducing the complexity of the dataset.

4.3.2 *Data Organisation*

Following the filtering, the next phase involves organizing the participants' data into specific directories based on their responses. This organizational step is crucial for efficiently managing and analysing data related to different Mini NYC-Q questions.

The organisational process involves:

1. **Loading the Filtered Data:** The filtered dataset is loaded from the CSV file into a data frame for easy manipulation.

2. **Preparing Directories:** A base directory structure is prepared to hold the sorted data files. This structure includes multiple subdirectories, each corresponding to a different Mini NYC-Q question item.

3. **Iterating Through Data Files:** The script iterates through the participant data files, extracting key information from the filenames to identify the participant and session.

4. **Matching Participant Data:** For each data file, the participant's ID is matched with their corresponding row in the filtered dataset to retrieve their responses to the Mini NYC-Q questions.
5. **Label Assignment:** Based on the participant's responses, a label ('pos' for positive or 'neg' for negative) is assigned to each filename.
6. **File Organization:** The files are then moved into the appropriate subdirectories based on the Mini NYC-Q question item they correspond to, ensuring they are sorted by both participant response and question item.

This comprehensive organizational process ensures that data files are systematically arranged, facilitating easier access and analysis. Each file is correctly labelled and placed in a directory corresponding to the specific Mini NYC-Q question item, streamlining the data management process for the music cognitive task study.

By filtering out unnecessary information and systematically organizing the data files, this approach enhances the efficiency and accuracy of the subsequent data analysis, making the process more manageable and less error-prone.

4.3.3 *Code Execution Flow*

The code execution starts with the importation of specialized libraries essential for its functionality. Among these, the 'glob' library is instrumental in expanding file paths, enabling the matching of pathnames and facilitating versatile handling of file paths for data retrieval from multiple files or directories. Another crucial library is 'mne', which is primarily designed for processing EEG (Electroencephalography) data. Also if this data was from another platform such as in our case being EEGLab. MNE streamlines EEG data manipulation, covering tasks such as loading, preprocessing, and feature extraction. Additionally, the 'shap' library, which stands for SHapley Additive exPlanations, provides a unified framework for interpreting the output of any machine learning model. It computes in our case the SHAP values to explain the output of a RandomForestClassifier, offering insights into feature importance and their impact on predictions.

Following the importation of libraries, the code defines several functions to streamline data loading, visualization, and result preservation. The dataset is then balanced by randomly oversampling the minority class to match the size of the majority class, preventing model bias towards the majority class. The results of said balancing and the distribution of positive and negative labels can be seen in Table 2.

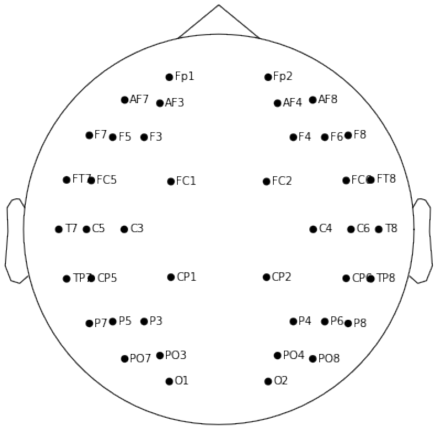

Figure 1: Music-Related Electrodes Transposed on the Scalp.

Using the 'load_data' function the code reads EEG data files, applies a bandpass filter, sets reference channels, and selects specific electrodes. As seen in Figure 1, electrodes chosen are electrodes that correlate to the brain regions which are active in music-related tasks Apicella et al. (2022). These include the following electrodes: Fp1, F7, Fp2, F8, F4, AF3, AF4, AF7, AF8, F3, F5, F6, FC1, FC2, FC5, FC6, FT7, FT8, C3, C4, C5, C6, T7, T8, TP7, TP8, CP1, CP2, CP5, CP6, P3, P4, P5, P6, P7, P8, PO3, PO4, PO7, PO8, O1, O2.

Fixed-length epochs are created from the continuous data, then converted into arrays for further analysis. Various statistical features (mean, standard deviation, peak-to-peak range, variance, minimum, maximum, etc.) are calculated for each epoch, capturing the characteristics of the EEG signals. The calculated features are concatenated into a single array for each epoch, forming the feature set for model training.

Group K-Fold cross-validation is used to ensure the data is split in a way that preserves the grouping of the epochs, preventing data leakage and ensuring robust evaluation. Classifiers are trained and evaluated using pipelines that include feature scaling and hyper parameter tuning with 'GridSearchCV'. The classifiers include algorithms: Logistic Regression, Naive Bayes, Random Forest, Support Vector Machines (SVM), Gradient Boosting Machines, and Neural Networks, are employed for predictive modelling tasks. Each algorithm has distinct characteristics suited to different types of data and classification problems.

Random Forest Classifier (RF):

- Description: An ensemble learning method that constructs multiple decision trees during training.

- Parameters Tuned:
  - 'n_estimators': Number of trees in the forest.
  - 'max_depth': Maximum depth of the trees.
  - Reasoning: Tuning these parameters controls model complexity and generalization capability. A grid search optimizes model performance.

Logistic Regression Classifier (LR):

- Description: A linear model suitable for binary classification tasks.
- Parameters Tuned:
  - 'C': Regularization strength.
  - Reasoning: Optimizing 'C' balances model simplicity and fitting the training data. Regularization prevents overfitting.

Naive Bayes (NB) Classifier:

- Description: A simple probabilistic classifier based on Bayes' theorem with strong independence assumptions between features.
- Parameters Tuned: None.
- Reasoning: Efficient with small datasets. Standard scaling maintains consistency across models.

Support Vector Machines (SVM) Classifier:

- Description: Effective for high-dimensional spaces, finding the hyperplane that best separates classes.
- Parameters Tuned:
  - 'C': Regularization strength.
  - 'kernel': Type of kernel function.
  - Reasoning: Adaptability to different datasets and decision boundaries is achieved by tuning 'C' and exploring different kernel functions.

Gradient Boosting Machines (GBM) Classifier:

- Description: Builds an ensemble of weak learners, sequentially improving performance.
- Parameters Tuned:

  - ‘learning_rate‘: Rate at which the model learns.
  - ‘max_depth‘: Maximum depth of the trees.
  - ‘n_estimators‘: Number of trees in the ensemble.
  - Reasoning: Tuning these parameters controls model complexity and performance. The model effectively captures underlying patterns in the data.

Neural Networks Classifier (NN):

- Description: A multi-layer perceptron with customizable architecture and activation functions.
- Parameters Tuned:
  - ‘hidden_layer_sizes‘: Sizes of the hidden layers.
  - ‘activation‘: Activation function.
  - Reasoning: Tuning architecture and activation functions allows the model to capture complex non-linear relationships in the data.

After model training the best scores from each model are compared and visualized using a horizontal bar plot, highlighting the most effective classifier. SHAP analysis with the goal of finding the optimal parameters according to the SHERPA method Sylvester et al. (2024) is conducted to explain predictions for the Random Forest classifier. This model was chosen due to it outperforming the others consistently while testing. SHAP values quantify the impact of each feature, particularly EEG electrodes, on the model’s output. This process provides insights into the relative importance of different electrodes in predicting target variables such as mental states.

To facilitate interpretation, the code includes a function named ‘plot_top_20_electrodes‘ to generate a bar plot visualizing the top 20 electrodes and their corresponding SHAP values. This visualization aids in understanding which electrodes exert the most substantial influence on the model’s predictions.

EEG data is now re-read and processed, focusing on the top 20 electrodes identified from the SHAP analysis on the Random Forest classifier. Similar to earlier steps, statistical features are extracted according to the data of the top electrodes, and all machine learning models are retrained and evaluated using these refined features. This is done so that we can see if preserving only the best performing electrodes impacts the accuracy of the models.

Finally, the code ensures the preservation of generated figures and results within designated directories. This practice safeguards analysis outcomes for subsequent scrutiny and interpretation, including creating the plot directory if absent, saving figures, and compiling results into a file denoted as 'results_filename'. This approach to result preservation enhances the reproducibility and transparency of the analysis process, enabling further examination and validation of the findings.

The entire process integrates data preprocessing, feature extraction, model training, evaluation, and interpretation, ensuring a comprehensive analysis pipeline for EEG data in the context of musical stimuli classification. By focusing on feature importance and refining the analysis to the most significant electrodes, the pipeline aims not only for high classification accuracy but also for meaningful insights into the underlying neural mechanisms. This approach combines advanced data processing techniques with robust machine learning practices, creating a powerful framework for EEG data analysis.

5 RESULTS

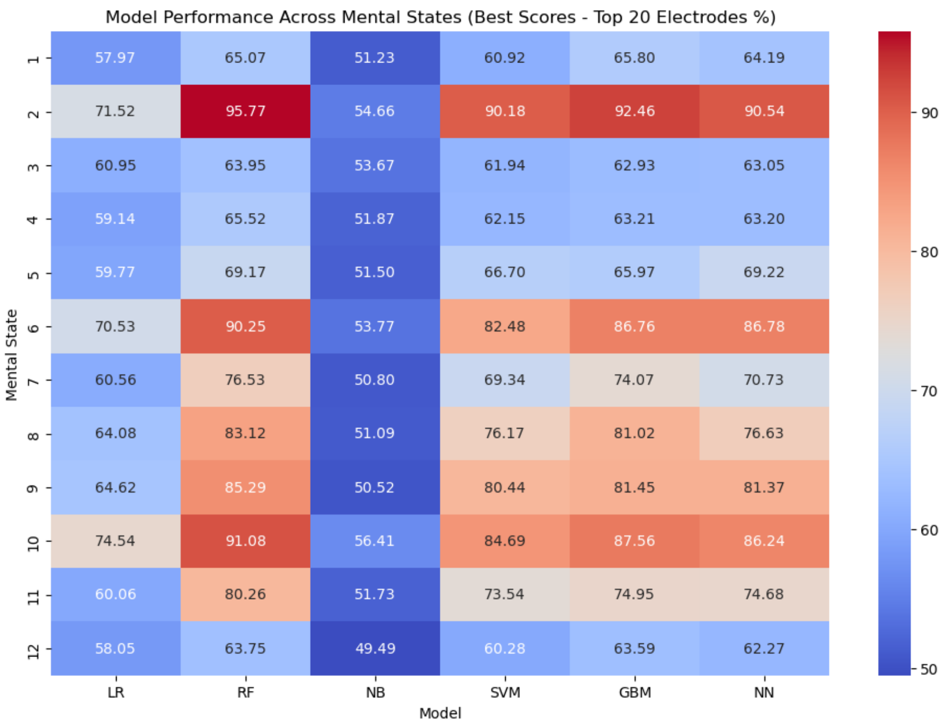

Figure 2: Comparison of Model Accuracy Across Different Mental States Using Top 20 Electrodes.

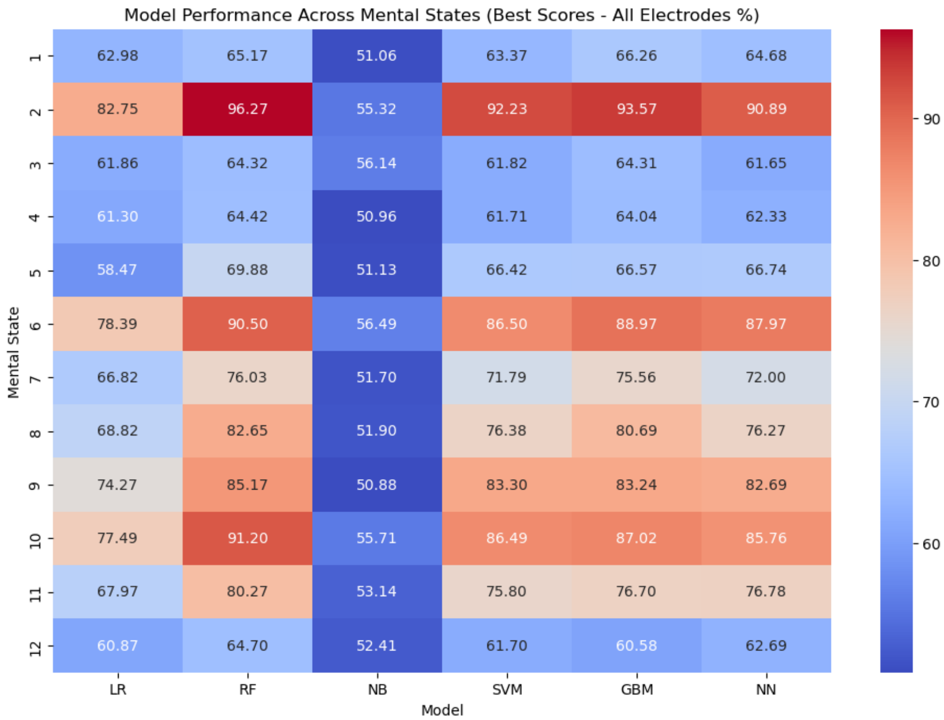

Figure 3: Comparison of Model Accuracy Across Different Mental States Using All Electrodes.

The overviews provided in figures 2 and 3 offer insights into the performance of various machine learning algorithms employed in classifying different mental states as assessed by the Mini NYC-Q. Figure 2 provides a detailed heat map of the classification results using all electrodes, showcasing the nuanced differences in accuracy across the spectrum of mental states. Meanwhile, figure 3 presents a similar heat map but focuses on results derived from the top 20 electrodes identified through the SHERPA method.

This observation is further substantiated by examining the exact accuracy figures presented in Appendix B (page 26). Notably, the table highlights significant performance disparities among the algorithms, with Naive Bayes consistently exhibiting the lowest accuracy across all mental states analysed. In contrast, Random Forest consistently demonstrates superior performance, achieving the highest accuracies across most mental states.

Across the diverse range of mental states analysed, classification accuracies generally fall within a moderate to high range, ranging approximately from 49.49% to 96.27%, depending on the dataset used and the specific mental state under consideration. These results underscore the variability

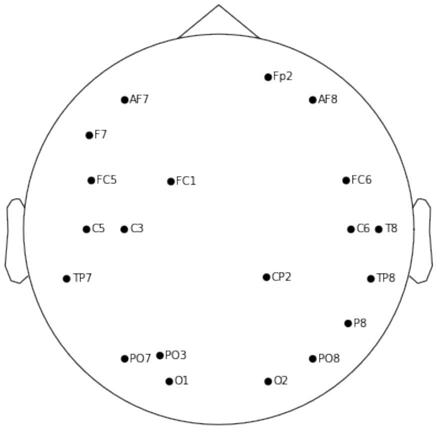

Figure 4: Music-Related Top 20 Electrodes Transposed on the Scalp Using Question 7 of the Mini Nyc-Q and the Random Forest Classifier.

in algorithmic performance and the challenge of accurately classifying mental states using machine learning approaches.

Interestingly, the classification accuracies remain largely consistent between the scenarios before and after electrode selection using the SHERPA method. This observation suggests that the selected electrodes effectively capture relevant neural information while minimizing redundant data and addressing privacy concerns. Such robust electrode selection methodologies hold promise for practical applications where data minimization and privacy preservation are critical considerations.

A detailed analysis of the data reveals variability in the top 20 electrodes associated with different mental states. These electrodes are crucial as they represent specific points on the scalp where neural activity is measured. Each electrode captures electrical signals generated by underlying neural processes, providing insights into brain activity related to different mental states.

While examining the dataset, several electrodes consistently appear among the top 20 across various mental states. These include electrodes such as F7, F4, F3, FC1, FC2, PO8, AF4, AF3, AF7, AF8, CP5, CP2, TP8, TP7, C3, C4, C5, C6, FT8, FT7, Fp1, and Fp2. Each of these electrodes corresponds to specific brain regions and neural networks involved in cognitive and emotional processing.

Interestingly, no single electrode consistently appears in the top 20 across all mental states analyzed. This variability suggests that different mental states may engage unique neural circuits or patterns of activity across the brain. For instance, while electrode F7 might be prominent in classifying one mental state, electrode TP8 might be more influential in another. Figure 4 illustrates what the top 20 electrodes of question 7 look like. Appendix A (page 26) shows the top 20 electrodes for each Mini NYC-Q.

6 DISCUSSION

The findings of this research reveal significant insights into EEG-based mental state classification. The results demonstrate that various machine learning algorithms, including Random Forest, Support Vector Machines (SVM), Gradient Boosting Machines, and Neural Networks, can achieve high accuracy in predicting mental states from EEG data. Specifically, Random Forest consistently outperformed other models, achieving the highest accuracies across most mental states examined. Interestingly, the classification accuracies remained largely consistent before and after the selection of the top 20 electrodes using the SHERPA method. This suggests that the selected electrodes effectively capture the relevant neural information while minimizing redundant data and addressing privacy concerns. Such robust electrode selection methodologies hold promise for practical applications where data minimization and privacy preservation are critical.

One surprising finding was that no single electrode consistently appeared in the top 20 across all mental states analysed. This variability implies that different mental states may engage unique neural circuits or patterns of activity across the brain. It could be speculated that this variability is due to the complex and dynamic nature of brain activity, which is influenced by a myriad of factors including individual differences and contextual elements.

6.1 *Limitations and Solutions*

Several limitations were acknowledged in this study:

1. Balancing of Data: The main limitation encountered was related to the balancing of the data. The use of random oversampling sometimes required the minority class to multiply itself multiple times to match the frequency of the majority class, leading to overfitting in some cases. A potential future solution could involve generating synthetic data through Radius-SMOTE, as mentioned by Wardayo et al. (2022).

2. Sample Size and Demographics: The dataset was initially limited to 60 participants, and after processing, it was reduced to just 57. This significantly restricted the number of positive or negative labels, leading to overfitting due to the heavy oversampling of certain questions from the Mini NYC-Q. The dataset predominantly included young adults aged 18-28. This demographic limitation could affect the generalizability of the

findings, as research has shown that aging can impact cognitive functions and mental states Peters (2006). Future studies should aim to include a more diverse sample in terms of age, ethnicity, and other demographic variables.

3. Data Quality and Preprocessing: While rigorous preprocessing steps were implemented, including re-referencing, filtering, and ICA for artefact removal, the potential for residual noise and artefacts remains. Advanced preprocessing techniques and higher-quality EEG equipment could enhance data quality.

6.2 *Contribution of the Study*

This research makes several key contributions:

1.Integration of Responsible AI Practices: By incorporating explainability, fairness, safety, and privacy into the model development process, this study sets a precedent for ethically responsible AI in the context of EEG data analysis.

2. Methodological Advancements: The use of SHERPA for electrode selection and SHAP for model interpretability provides a framework for future studies aiming to balance accuracy with ethical considerations.

3. Practical Applications: The findings have potential applications in various fields, including healthcare, education, and human-computer interaction, where understanding and predicting mental states can significantly enhance user experiences and outcomes.

The results of this study provide critical insights into the feasibility and ethical considerations of using machine learning algorithms for EEG-based mental state classification. Additionally, the successful application of SHERPA through SHAP methodologies underscores the importance of interpretability and data minimization in developing responsible AI systems.

By addressing the challenges and limitations encountered, future research can build upon these findings to enhance the accuracy, fairness, and ethical integrity of EEG-based mental state classification models. This study, therefore, contributes to the advancement of both cognitive science and responsible AI, paving the way for more trustworthy and effective AI applications in neuroscience.

7 CONCLUSION

This study was guided by the overarching research question: "How can machine learning algorithms be effectively utilized to predict mental states based on EEG data while ensuring responsible AI practices are upheld throughout the model development and evaluation process?" To explore this, three sub-questions were addressed:

1. What ethical considerations need to be taken into account when working with EEG data for mental state classification?

2. How can model interpretability techniques such as SHAP contribute to understanding the decision-making process of the classification model?

3. How can the classification model be fine-tuned to balance accuracy and ethical considerations, ensuring responsible AI practices are maintained throughout the project?

The study successfully demonstrated that machine learning algorithms, particularly the Random Forest model, can effectively predict mental states from EEG data with high accuracy. The Random Forest model consistently outperformed other models, achieving superior classification accuracy across various mental states. This aligns with previous research by Liu et al. 2013 and Hu et al. 2018, which also found high accuracies in mental state classification using similar techniques.

In terms of ethical considerations, the integration of responsible AI practices was a significant advancement. The use of SHAP for model interpretability provided insights into the decision-making process of the classification models, ensuring transparency. The SHERPA method through the SHAP analysis for electrode selection helped to minimize data collection, addressing privacy concerns while maintaining model performance. This approach underscores the importance of ethical practices in AI, as highlighted by Coin et al. (2020) and Molnar (2023).

The findings of this study have several important implications:

1. Ethical AI Development: The successful integration of ethical considerations such as transparency, fairness, safety, and privacy in the model development process provides a blueprint for future AI research and applications, particularly those involving sensitive data like EEG recordings.

2. Model Interpretability and Data Minimization: The methodologies employed in this study, including SHERPA through SHAP, demonstrates an effective way to achieve model interpretability and data minimization. These techniques can be applied to other domains where similar ethical considerations are paramount.

3. Practical Applications: The high accuracy of the Random Forest model and the effective electrode selection method have potential applications in various fields. For example, in healthcare, these models can be used for non-invasive mental health monitoring and diagnosis. In education, they can help in assessing student engagement and cognitive load, leading to more personalized learning experiences.

In conclusion, this study has successfully demonstrated the feasibility and ethical considerations of using machine learning for EEG-based mental state classification. By integrating responsible AI practices, it provides a solid foundation for future research aimed at developing trustworthy and effective AI applications in neuroscience and beyond.

REFERENCES

Acı, , Kaya, M., & Mishchenko, Y. (2019). Distinguishing mental attention states of humans via an eeg-based passive bci using machine learning methods. *Expert Systems With Applications*, *134*, 153–166. Retrieved from https://doi.org/10.1016/j.eswa.2019.05.057 doi: 10.1016/j.eswa.2019.05.057

Anwer, T. (2021, 12 1). *Eeg read signal, process and machine learning classification using python.* YouTube. Retrieved from https://youtu.be/_BdBJOOqMes?si=Ynxa8UH7F0JfTvYl

Apicella, A., Arpaïa, P., Frosolone, M., Improta, G., Moccaldi, N., & Pollastro, A. (2022). Eeg-based measurement system for monitoring student engagement in learning 4.0. *Scientific Reports*, *12*(1). Retrieved from https://doi.org/10.1038/s41598-022-09578-y doi: 10.1038/s41598-022-09578-y

Coin, A., Mulder, M., & Dubljević, V. (2020). Ethical aspects of bci technology: What is the state of the art? *Philosophies*, *5*(4), 31. Retrieved from https://doi.org/10.3390/philosophies5040031 doi: 10.3390/philosophies5040031

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. Retrieved from https://sccn.ucsd.edu/eeglab/index.php

European Union. (2016). *General data protection regulation (GDPR) (EU) 2016/679.* Official Journal of the European Union. Retrieved from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679

Gorgolewski, K. J., Lurie, D. J., Urchs, S., Kipping, J., Craddock, R. C., Milham, M. P., ... Smallwood, J. (2014). A correspondence between individual differences in the brain's intrinsic functional architecture and the content and form of self-generated thoughts. *PLOS ONE*, *9*(5), e97176. Retrieved from https://doi.org/10.1371/journal.pone.0097176 doi: 10.1371/journal.pone.0097176

hamseelmi199. (2024). *CSAI_Bachelor_Thesis.* GitHub. Retrieved from https://github.com/hamseelmi199/CSAI_Bachelor_Thesis

Hu, B., Li, X., Sun, S., & Ratcliffe, M. (2018). Attention recognition in eeg-based affective learning research using cfs+knn algorithm. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *15*(1), 38–45.

Retrieved from https://doi.org/10.1109/tcbb.2016.2616395 doi: 10.1109/tcbb.2016.2616395

Liu, N., Chiang, C.-Y., & Chu, H.-K. (2013). Recognizing the degree of human attention using eeg signals from mobile sensors. *Sensors*, *13*(8), 10273–10286. Retrieved from https://doi.org/10.3390/s130810273 doi: 10.3390/s130810273

Maghsoudi, A., & Shalbaf, A. (2021). Mental arithmetic task recognition using effective connectivity and hierarchical feature selection from eeg signals. *Basic and Clinical Neuroscience*, *12*(6), 817–826. Retrieved from https://doi.org/10.32598/bcn.2021.2034.1 doi: 10.32598/bcn.2021.2034.1

Molnar, C. (2023). *Interpretable machine learning*. Retrieved from https://christophm.github.io/interpretable-ml-book/

OpenAI. (2023). *ChatGPT (3.5 version).* Large language model. Retrieved from https://chat.openai.com/chat

Peng, C.-J., Chen, Y.-C., Chen, C., Chen, S.-J., Cagneau, B., & Chassagne, L. (2019). An eeg-based attentiveness recognition system using hilbert–huang transform and support vector machine. *Journal of Medical and Biological Engineering*, *40*(2), 230–238. Retrieved from https://doi.org/10.1007/s40846-019-00500-y doi: 10.1007/s40846-019-00500-y

Peters, R. (2006). Ageing and the brain. *Postgraduate Medical Journal*, *82*(964), 84–88. Retrieved from https://doi.org/10.1136/pgmj.2005.036665 doi: 10.1136/pgmj.2005.036665

Ryali, S., Zhang, Y., de los Angeles, C., Supekar, K., & Menon, V. (2024). Deep learning models reveal replicable, generalizable, and behaviorally relevant sex differences in human functional brain organization. *Proceedings of the National Academy of Sciences*, *121*(9), e2310012121. Retrieved from https://doi.org/10.1073/pnas.2310012121 doi: 10.1073/pnas.2310012121

Scrivener, C. L., & Reader, A. T. (2022). Variability of eeg electrode positions and their underlying brain regions: Visualizing gel artifacts from a simultaneous eeg-fmri dataset. *Brain and Behavior*, *12*(2), e2476. Retrieved from https://doi.org/10.1002/brb3.2476 doi: 10.1002/brb3.2476

Suhail, T. A., Indiradevi, K., Suhara, E. M., Poovathinal, S. A., & Ayyappan, A. (2022). Distinguishing cognitive states using electroencephalography local activation and functional-connectivity patterns. *Biomedical Signal Processing and Control*, *77*, 103742. Retrieved from https://doi.org/10.1016/j.bspc.2022.103742 doi: 10.1016/j.bspc.2022.103742

Sylvester, S., Sagehorn, M., Gruber, T., Atzmueller, M., & Schöne, B. (2024).

Shap value-based erp analysis (sherpa): Increasing the sensitivity of eeg signals with explainable ai methods. *Behavior Research Methods*. Retrieved from https://doi.org/10.3758/s13428-023-02335-7 (Advance online publication) doi: 10.3758/s13428-023-02335-7

Toa, C. K., Sim, K. S., & Tan, S. C. (2021). Electroencephalogram-based attention level classification using convolution attention memory neural network. *ing and the Brain." Postgraduate Medical Journal 82, no. 964*. Retrieved from https://doi.org/10.1136/pgmj.2005.036665 doi: 10.1136/pgmj.2005.036665

Wang, Y., Duan, W., Dong, D., Ding, L., & Lei, X. (2022a). *A test-retest resting and cognitive state eeg dataset [data set].* OpenNeuro. Retrieved from https://doi.org/10.18112/openneuro.ds004148.v1.0.1

Wang, Y., Duan, W., Dong, D., Ding, L., & Lei, X. (2022b). A test-retest resting, and cognitive state eeg dataset during multiple subject-driven states. *Scientific Data*, *9*(1). Retrieved from https://doi.org/10.1038/s41597-022-01607-9 doi: 10.1038/s41597-022-01607-9

APPENDIX A

Mental State	Top 20 Electrodes
1	O1, TP8, FC1, CP2, P8, PO7, TP7, C6, O2, AF8, T8, FC5, C5, PO8, C3, PO3, FC6, Fp2, AF7, F7
2	P6, P7, F4, FC1, P5, AF4, CP6, F6, FC2, C5, T8, O1, CP2, T7, F7, AF7, AF8, TP7, P8, F3
3	FC1, P7, P6, CP1, F8, TP7, CP5, F3, AF3, P3, FC2, T7, O2, CP6, PO8, CP2, FC5, C5, F6, P8
4	F4, P7, C6, F6, AF7, CP6, F7, AF4, O2, C5, AF8, P3, FC2, FC6, PO8, PO4, TP8, C3, P8, FC5
5	FC1, FC2, O1, FT8, CP2, TP8, C4, F3, CP6, O2, P8, CP1, Fp2, FC6, Fp1, CP5, AF3, AF8, FC5, T8
6	FC1, P7, O1, PO7, FC2, AF8, C6, PO8, C4, O2, CP2, FT8, FC5, FC6, C3, Fp1, AF7, PO4, TP8, CP1
7	P6, P4, FC2, F4, P7, AF4, P8, TP8, O2, CP6, TP7, PO8, O1, PO7, CP1, FC6, F3, F5, AF8, Fp1
8	P4, TP8, F5, P8, F3, AF3, CP1, O1, P5, CP5, F6, O2, CP6, TP7, AF4, C3, Fp2, FC5, FC2, FT8
9	FC1, O1, CP1, PO8, TP8, P7, O2, CP2, P8, Fp1, Fp2, C3, C6, TP7, PO4, AF8, PO3, C4, FC5, CP5
10	P7, P6, F3, AF3, F4, T8, P4, FC1, TP7, P3, AF4, CP6, C4, T7, FT8, PO4, PO7, PO3, C3, CP2
11	FC1, F8, CP6, CP5, AF8, TP8, C3, F5, AF7, P8, CP2, FT8, PO8, AF3, F7, O1, Fp1, PO4, O2, FT7
12	CP2, O1, P7, P5, CP1, FC1, P3, TP7, F6, PO8, FC2, PO3, P8, C3, TP8, Fp1, AF4, Fp2, PO7, PO4

Table 3: Top 20 Electrodes for Each Mental State.

APPENDIX B

Mental State	LR	RF	NB	SVM	GBM	NN
Best Scores (All Electrodes)						
1	62.98%	65.17%	51.06%	63.37%	66.26%	64.68%
2	82.75%	96.27%	55.32%	92.23%	93.57%	90.89%
3	61.86%	64.32%	56.14%	61.82%	64.31%	61.65%
4	61.30%	64.42%	50.96%	61.71%	64.04%	62.33%
5	58.47%	69.88%	51.13%	66.42%	66.57%	66.74%
6	78.39%	90.50%	56.49%	86.50%	88.97%	87.97%
7	66.82%	76.03%	51.70%	71.79%	75.56%	72.00%
8	68.82%	82.65%	51.90%	76.38%	80.69%	76.27%
9	74.27%	85.17%	50.88%	83.30%	83.24%	82.69%
10	77.49%	91.20%	55.71%	86.49%	87.02%	85.76%
11	67.97%	80.27%	53.14%	75.80%	76.70%	76.78%
12	60.87%	64.70%	52.41%	61.70%	60.58%	62.69%
Best Scores (Top 20 Electrodes)						
1	57.97%	65.07%	51.22%	60.92%	65.80%	64.19%
2	71.52%	95.77%	54.66%	90.18%	92.46%	90.54%
3	60.95%	63.95%	53.67%	61.94%	62.93%	63.05%
4	59.14%	65.52%	51.87%	62.15%	63.21%	63.20%
5	59.77%	69.17%	51.50%	66.70%	65.97%	69.22%
6	70.53%	90.25%	53.77%	82.48%	86.76%	86.78%
7	60.56%	76.53%	50.80%	69.34%	74.07%	70.73%
8	64.08%	83.12%	51.09%	76.17%	81.02%	76.63%
9	64.62%	85.29%	50.52%	80.44%	81.45%	81.37%
10	74.54%	91.08%	56.41%	84.69%	87.56%	86.24%
11	60.06%	80.26%	51.73%	73.54%	74.95%	74.68%
12	58.05%	63.75%	49.49%	60.28%	63.59%	62.27%

Table 4: Best Scores for Each Model With All Electrodes and Top 20 Electrodes.