EEG Thinking1 Datasets: Think-Count-Recall (TCR) and Read-Write-Type (RWT)

Xiaodong Qu, Peiyan Liu
Department of Computer Science
Brandeis University
Waltham, MA 02453
xiqu,peiyanliu@brandeis.edu

Abstract

EEG-based Brain-Computer Interfaces (BCI) have been widely used in clinical and non-clinical research. In this paper, we present a framework to collect a large amount of EEG data with easy-to-use experiment setup, using non-invasive, wireless, and affordable hardware. Interpretable feedback generated by benchmark machine learning algorithms have been provided to the researchers and end-users. Two existing datasets are used as case studies for the framework: Read-Write-Type (RWT) and Think-Count-Recall (TCR). The goal is to inspire new machine learning approaches for decoding behavior from large-scale EEG data. The framework of experimental design, data collection, data analysis, feedback generation, and community building could pave the way towards a future when everyone can easily use BCI systems every day, similar to smartphones nowadays.

1 Introduction

Neural interfaces are becoming of increasing interest to industry and having large available datasets could be useful for students and researchers to tease out signals from noisy data. Brain Computer Interfaces (BCI) have been widely used for both clinical and non-clinical applications (Lotte et al. [2018a], Craik et al. [2019]), such as diagnosis of abnormal states, evaluating the effect of the treatments, helping patients with motor disabilities to move a mouse or to control a motorized wheelchair, mental workload, seizure detection, motor imagery tasks (Devlaminck et al. [2010]), BCI based games (Coyle et al. [2013]) and passive BCI. Previous research has reviewed existing datasets in the BCI field, such as Schalk et al. [2004], Lotte et al. [2007], Zhang et al. [2020], Roy et al. [2019], Miller [2019], Kaya et al. [2018], most of the datasets mentioned are collected in research labs or clinical settings with expensive medical equipment and time-consuming setup procedure, under the supervision of clinical professionals. The data collection framework we proposed allows non-expert participants to run the experiment by themselves at home, whenever they have a small amount of time, such as twenty minutes. The visual feedback generated by benchmark machine learning algorithms could help them to perform better in the future sessions.

Considering classic datasets in other domains, such as ImageNet for image classification, or MNIST for handwritten digit recognition, more data can be generated directly from the non-expert end users, and more general patterns could be recognized based on such large scale data. With the motivation to gather EEG data with a cheaper, easier and faster approach, we designed a pilot study towards building a large-scale EEG data set, for multi-class classification of user-centered tasks, generated by non-expert end-users. Results of classification with the proposed new data and machine models show a reasonable accuracy (70% to the random 20%), indicating the potential of this framework.

Submitted to the 35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks. Do not distribute.

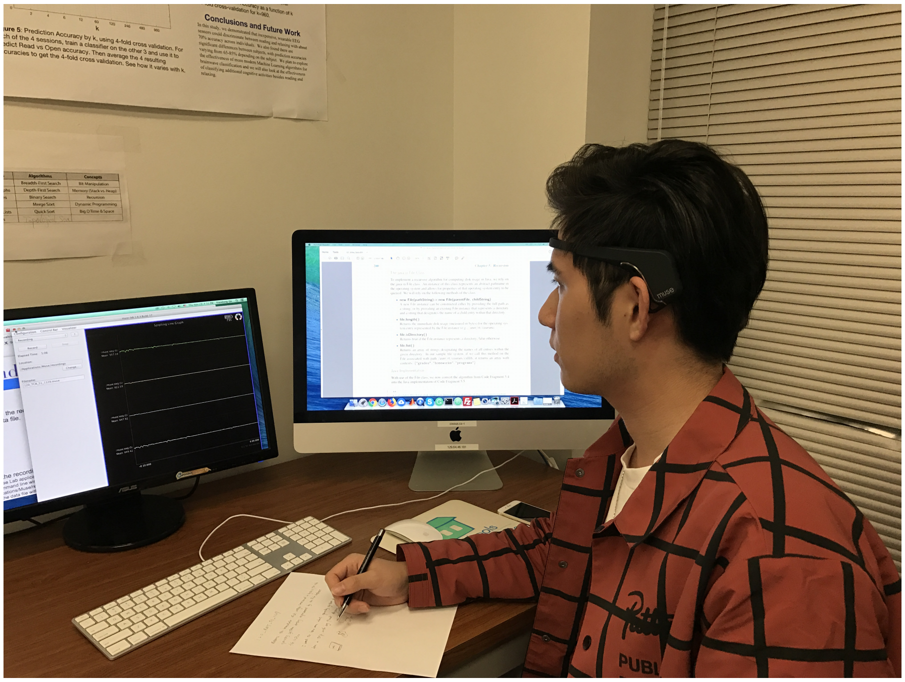

Figure 1: A researcher demonstrating the task "Write", wearing EEG headset.

This is an ongoing project and this 'Thinking1' repository currently has four datasets: Read-Write-Type (RWT, Qu et al. [2020b]), Think-Count-Recall (TCR, Qu et al. [2020a]), Python-Math (Qu et al. [2018b]), and GRE-Relax (Qu et al. [2018a]). In this paper, we use the two recent experiments, RWT and TCR, as examples to explain the approach. Details about the data collection, including how the subjects were recruited under IRB requirements, how long each session was, what kind of visualized feedback is provided to the subjects, how many EEG sessions were recorded, data cleaning, feature extraction, and research from benchmark algorithms, are in our previous papers, we also attached an updated version in the appendix section of this paper to allow readers to replicate these experiments.

1.1 Read-Write-Type (RWT)

Previous studies (Bird et al. [2018], Qu et al. [2018a]) demonstrated that EEG signals could successfully distinguish several kinds of cognitive tasks. Such as programming in Python vs. solving Math problems; solving Math problems (GRE) vs. solving Reading problems (GRE). These experiments focused on distinguishing different cognitive tasks, but not on whether different communication modes may also have a distinguishable impact on EEG patterns. The experiment RWT (Qu et al. [2020b]) in this data set was designed to test the hypothesis of whether AI based EEG markers could distinguish both between two modes of communication: typing vs. writing, and between three cognitive states: reading vs. copying vs. answering. The five tasks are described in Figure 2.

1.2 Think-Count-Recall (TCR)

Other studies (Lotte et al. [2018a], Lotte [2015], Bird et al. [2018], Qu et al. [2018a], Craik et al. [2019]) demonstrated that EEG classification was successfully used to distinguish multiple cognitive tasks. In the TCR experiment (Qu et al. [2020a]), we designed these five user-centered tasks as shown in Figure 3, abbreviated them as Think (T), Count (C), Recall (R), Breathe (B) and Draw (D). The task selection is motivated by human memory experiments such as Kahana et al. [2018].

2 Methods

Such datasets are suitable for machine learning due to its high dimensional and noisy nature, similar to image recognition problems. There is great potential to provide higher accuracy and more interpretable feedback to both researchers and end-users. For example, in each data point of 1/10 second, the raw EEG data is a 4 x 5 matrix, which represents four electrodes and five frequency bands. Such twenty-dimensional data performs well enough (compare to 64 or 128 electrodes medical devices) when applied to mainstream EEG-related machine learning or deep learning algorithms.

Each session of these experiments are reproducible with twenty minutes of effort for non-experienced end-users. These human-in-the-loop experimental designs motivated by (LaRocco et al. [2020], Lotte et al. [2018b]), have several advantages. First, the tasks are selected more from the end-users, less from the researchers, similar to the smartphone usage situation now. Secondly, the role of the EEG coach can make the end-user experience much better. Last but not the least, easy-to-understand user feedback could be helpful for the end-user to reduce the noise and focus more on the designed tasks. More details in the previous papers and the appendix section of this paper.

2.1 Experimental Design

The experimental design is easy to adapt, and the three hundred dollars or less wireless hardware, as mentioned in Ienca et al. [2018], makes it affordable to a broader audience. For example, our research lab has expanded the experiments from just targeting less than twenty students, to a community of more than one hundred students, each of them starts with little or no computer science or neuroscience background, and usually, after at most two to three twenty-minute sessions, they can learn to how to control the noise level, and achieve the desired experimental goals with high accuracy.

The sensor hardware research and development have grown rapidly recently (Kübler et al. [2014], Tabar and Halici [2016]), so does the trend of making it more affordable to the non-expert users. After comparing several options, such as devices mentioned in Ienca et al. [2018], we chose the Muse Headset for our experiments, with an affordable price of less than three hundred for each wireless headset. For the design of the tasks, previous research has shown deep learning works well in emotion recognition, motor imagery, mental workload, and seizure detection areas (Craik et al. [2019]), we tried learning, motor-imagery tasks, sleep, and entertainment tasks. In this study, we focused on the learning related tasks college students perform often in their daily lives.

2.2 Data collection

Data was collected in non-clinical settings, partly in the reserved classrooms or conference rooms in the universities, partly at the participants' home. The size of the data usually is 15 to 20 subjects, five to six sessions for each subject, each sessions varies from five minutes to twenty minutes. For example, the TCR (16 subjects) and RWT (14 subjects) experiment each includes six sessions, each session is five minutes long. Comparing with existing experiments on cognitive tasks mentioned in Craik et al. [2019], Gabard-Durnam et al. [2018], Roy et al. [2019], Pernet et al. [2019], our experimental design and data collection is easier, cheaper and faster. With twenty-minute training, most participants can generate hours of EEG recording data at home with interpretable feedback.

The non-invasive, wireless EEG headset usually needs a training session to reduce the noise level. The role of EEG coach was created to smooth the learning curve for first time end-users. The end-users and EEG coaches are fairly compensated under the IRB requirement. More details such as

Read (R) Subjects were asked to read a PDF file displayed on the monitor silently, the PDF file is a computer science textbook on Data Structures (Sierra and Bates [2003]).

Write Copy (WC) Subjects wrote on a blank white paper with a pen, copying the text from the same textbook PDF file display on the monitor. As shown in Figure 1.

Write Answer (WA) Subjects wrote an essay using a pen on a blank paper, answering the question: 'Why did you choose your major?'

Type Copy (TC) Subjects copied text from the same textbook PDF file, into a text entry box on the screen, by typing on a keyboard.

Type Answer (TA) Subjects typed their answers to the question 'What is your academic plan for this semester?' into a text entry box on the screen.

Figure 2: Tasks in experiment Read-Write-Type (RWT).

Think (T) Subjects were asked to think of several (six, seven, eight) random objects, these objects need to be independent of each other. For example, (Sun, Fish, Flower, Table, Student, Car), is a valid set, but (computer, keyboard, monitor, speaker, phone, TV) is not a valid set.

Count (C) Subjects counted numbers aloud, from 200 towards 0, each time subtracting by 7, e.g. 200, 193, 186, 179, with eyes open, eyes and jaws movement minimized.

Recall (R) Subject recalled the objects they had typed in the Think (T) task, in the correct order, if possible, and entered them in a similar text entry box with a keyboard.

Breathe (B) Subjects were instructed to breathe deeply with their eyes open. They were asked NOT to think about any other tasks in this experiment, or anything else except their breath.

Draw (D) Subjects were asked to draw the objects they thought about in the earlier task Think (T), with a pen, on a blank A4 paper. The objects text they just entered in T was displayed on the monitor, so they did not need to recall, just focus on drawing.

Figure 3: Tasks in experiment Think-Count-Recall (TCR).

IRB approval and instructions given to the participants are included in the appendix section of this paper. Each headset was connected to a mainstream personal computer through Bluetooth. We use the software package that comes with the EEG headset (Muse-io and MuseLab) to record the raw EEG data to the computers. Then the data was processed and Analyze using machine learning and deep learning algorithms. The visualized feedback is provided to the end-users, EEG coaches, and researchers to improve the next round of data collection.

Before the experiment, the EEG coach helps the end-users to understand the IRB requirements and make sure they sign the informed consent forms, then explain in detail to the end-users what they are expected to see and to do during each step. During each session of the experiments, the EEG coach leads the end-users to the experiment website to fill out the pre-experiment survey, then helps the end-users to connect the EEG headsets and conduct a test recording for one minute before the official EEG recording starts, A time-boxed online survey style guide was then used to give the end-user step-by-step prompt during the experiment, the EEG coach is there for any possible questions. After the experiment, The EEG coach makes sure the end-user fill out the post-experiment survey and help them better understand the visual feedback. Also, the EEG coach keeps track of the notes for the entire process and communicates with the researchers regularly to deal with pop-up issues and maintain a frequent-asked-question (FAQ) list.

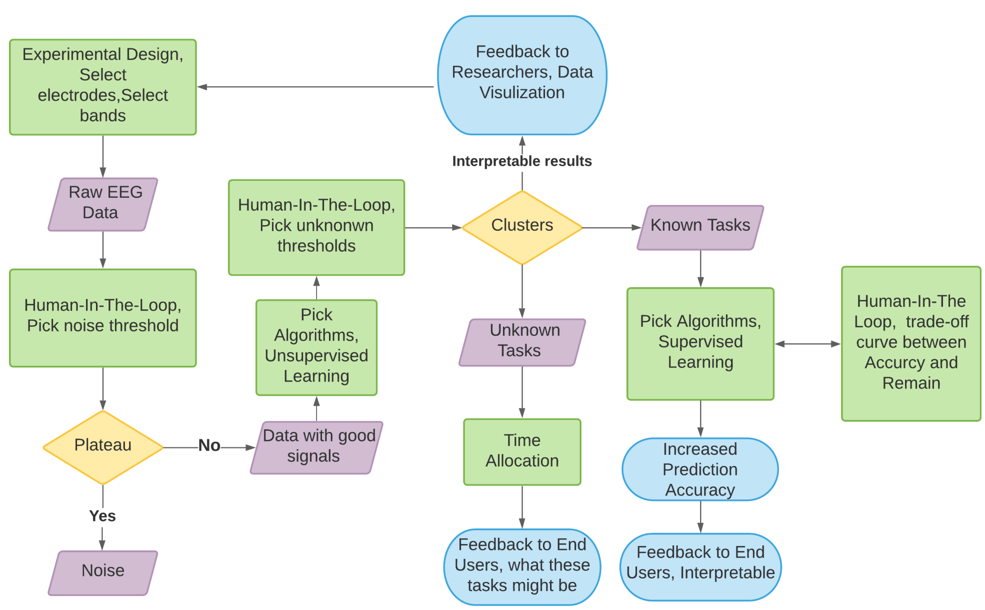

Figure 4: Data analysis framework.

2.3 Data analysis framework

New ways to analyze EEG data have been developed recently, such as Chevalier et al. [2020], Roc et al. [2020], Sabbagh et al. [2019], Tu et al. [2019]. In our experiments, as Figure 4 shows, the raw EEG data were first visualized to allow researchers to define the threshold to remove the noise, here we used the plateau threshold to determine whether a certain time window of signals is considered noise. Then we first used unsupervised learning algorithms, such as K-means, to cluster the data points, then we visualized the clustering result and made it interpretable to the researchers and end-user. For the designed tasks, we then used supervised learning, such as Random Forest or Long Short-Term Memory (LSTM) to predict the tasks. For the unknown tasks, we put a marker on the visual feedback to the end-user to ask what may happen during that time period. More details about these steps are in the appendix section of this paper.

By Human-In-The-Loop (HITL) it means the researcher, EEG coaches, and end-users and making data-driven decisions based on visualized feedback. The researchers bring together the existing knowledge about this experiment and help the EEG coaches and end-users to perform better in the next session. The EEG coaches guided the end-users to master how to use EEG hardware and software necessary to perform the designed tasks. The end-users learn from the researchers and the EEG coaches to use the EEG-based on BCI in their daily life as the experiments designed, and provide valuable feedback to the researchers to develop new and improved experiments.

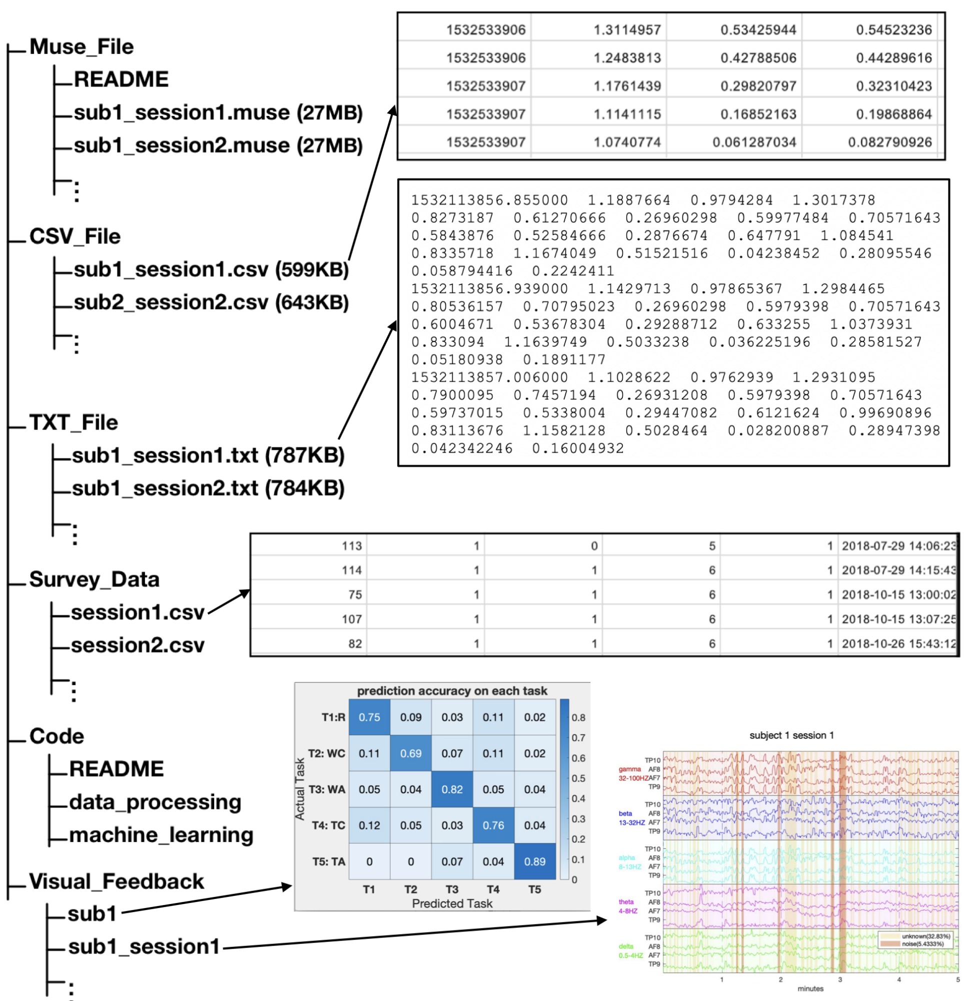

Figure 5: File structure of Our dataset

2.4 Data format

Pernet et al. [2019], and Nichols et al. [2017] have presented several recommended practices about EEG data formats and sharing. Our data set, as Figure 5 shows, consists of the original MUSE files, and CSV files, TXT files after pre-processing. Also, the metadata collected through Qualtrics online survey system has been included, as well as the code has been implemented for this dataset. For example, for the Read-Write-Type (RWT) experiments, for each subject, each five-minute session, there is a MUSE data file size of around 27M, and after pre-processing, the MUSE file is converted to a CSV or TXT file for further analysis, with a much smaller size of about 700K. Then there are folders of suvery metadata and related code.

2.5 Machine Learning applications

In this paper, we introduce a machine learning benchmark for predicting the task humans are engaged in from the EEG. We presented what machine learning and deep learning algorithms have been applied to these datasets, and suggest several recommended practices for these datasets.

For the pre-processing part, data visualization is helpful for noise detection. For the multi-class classification, ensemble methods, such as random forest, and Recurrent Neural Network (RNN), such as Long Short-Term Memory (LSTM) consistently outperformed other classifiers (Qu et al. [2020b]), we suggest using them as benchmark algorithms. Building on top of that, we proposed our algorithm, Time-Continuity-Voting (TCV, Qu et al. [2020a]), which achieved the highest prediction accuracy for these datasets. More details are in the appendix section of this paper.

2.6 Community Forum and further support

We established an online forum for the community who works on these datasets, including researchers, EEG coaches, end-users, and clinical professionals. Due to our IRB requirements, this forum is invitation-only at this time. Through our BCI forum, we connected to three computer science labs, two neuroscience labs, two clinical research labs, and two hospitals during the last three years, as well as get more than a hundred undergraduate students involved as experiment participants, eight of them later became EEG coaches.

We held discussions on how to improve EEG experimental design and dataset development. Further support on how to explore the potential of such an EEG-based BCI system is encouraged based on community members' availability. Also, we are presenting these research papers and this forum to more college students in the computer science and Neuroscience courses we lectured each semester.

Participating in the existing BCI community and bridging our own small EEG-based BCI community to a broad network is also an important direction.

2.7 Availability and Ethical considerations

To make sure these datasets would be used ethically and responsibly, we adapted several recommended practices of sharing BCI data, such as Gabard-Durnam et al. [2018], Pernet et al. [2019]. According to our IRB requirement, these data sets are available upon written request, we review the request to make sure it is coming from a reputable research institution and the requester is willing to sign a Non-Disclosure Agreement. Previous studies have reviewed freely available EEG datasets, such as Zhang et al. [2020], Roy et al. [2019], Miller [2019], Kaya et al. [2018], Craik et al. [2019], we are amending our IRB to find acceptable ways of data anonymization to share it more freely.

3 Results

We develop feedback for different user roles. For example, the figures that compare different end-users or different machine learning algorithms are more for the researchers, optional for the end-users. Here are some sample feedback figures we provide to our researchers, EEG coaches and end-users.

3.1 For Researchers

For cross-subject comparison, as Figure 6 shows, although there are individual differences, the task prediction accuracy is reasonably high. Together with Figure 7, (both figures X-axis is subject id ordered by prediction accuracy), we observed the noise and unknown tasks vary across different

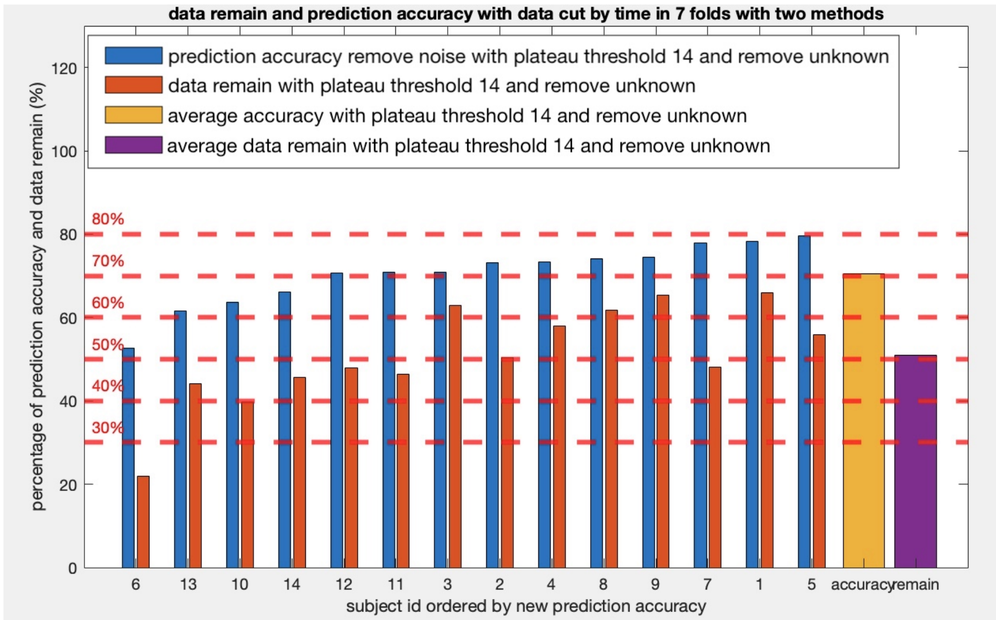

Figure 6: Experiment RWT: task prediction accuracy and data remain.

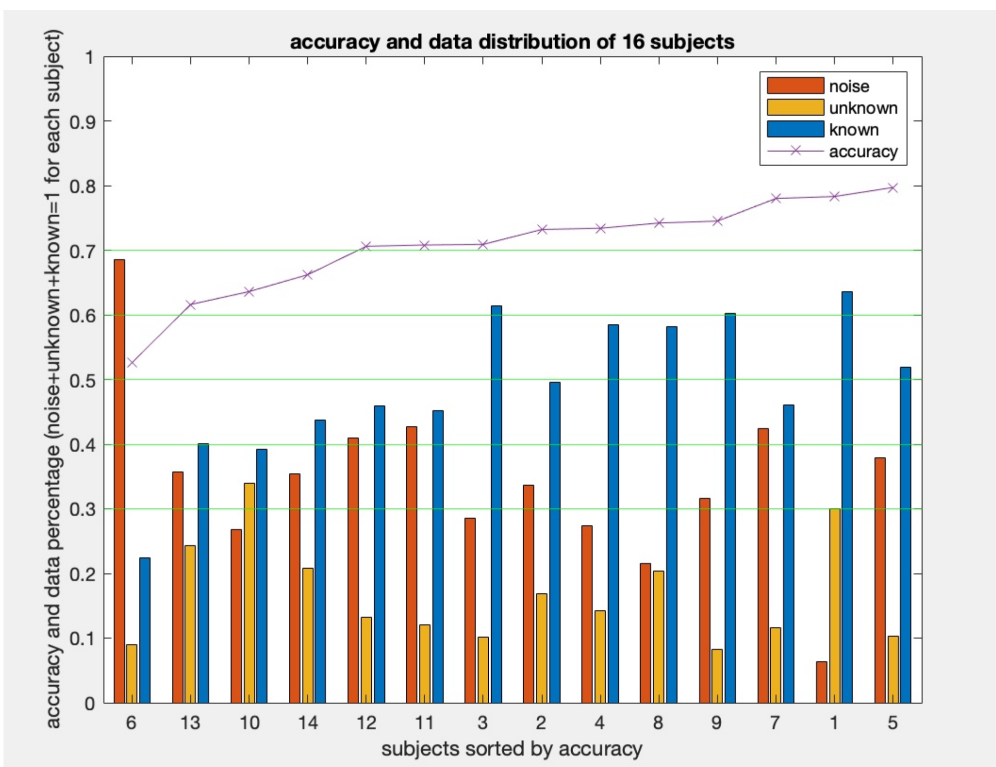

Figure 7: Experiment RWT: noise, unknown, and known tasks percentage.

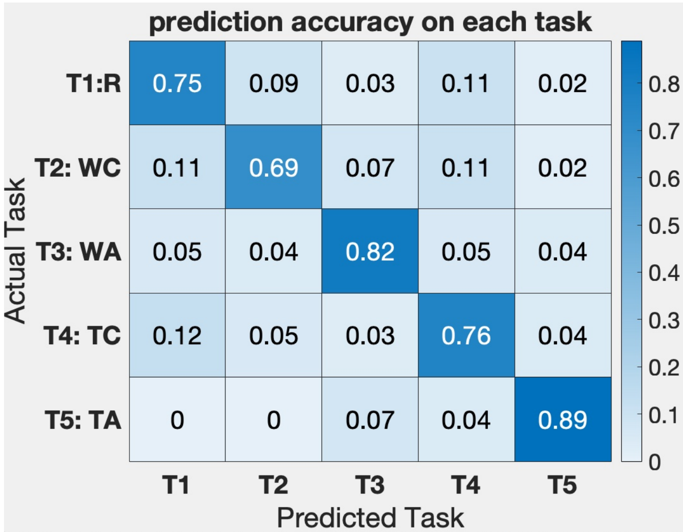

Figure 8: Diagonal Accuracy

subjects. Thus end-user training is necessary for better controlling the noise and unknown tasks. The role of EEG coach is created for this purpose.

3.2 For End-Users

Figure 8 shows for subject one in experiment RWT, how the accuracy of each task is predicted over all six sessions. This feedback may guide the further task selections. Each individual has a unique task set that is easy to be recognized with this EEG-based BCI experimental design and data collection framework. Thus it has the potential to be used as personal EEG fingerprint. Figure 9 shows the noise

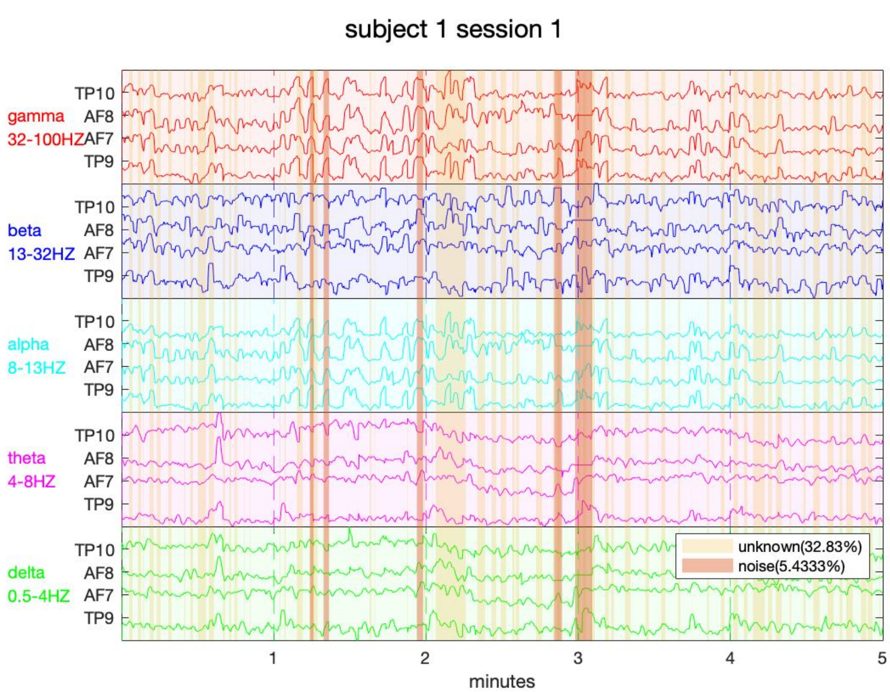

Figure 9: Noise, unknown, and known tasks, experiment RWT: subject one, session one.

and unknown task locations in each session, such feedback is helpful for the end-user to reflect what happened around a certain time spot.

4 Discussion

The main goal for this paper is to provide a framework of experimental design and data collection to gather EEG data through cheap means and non-expert participants. Intepretable feedback generate by benchmark machine learning algorithms can speed up this process. Comparing with the traditional data collection methods, as mentioned in Craik et al. [2019], Gabard-Durnam et al. [2018], Roy et al. [2019], Pernet et al. [2019], our approach is faster and cheaper to gather more EEG data with non-expert participants. Our efforts are made toward the future of everyone can use EEG-based BCI in their daily life, similar to the current everyday usage of smartphones. Although the limitation of sensory accuracy will remain for a while, the research related to the non-invasive BCI shows a growing potential to reach out to non-expert end-users.

The datasets we present are an early exploration of how to map the healthy subjects' daily activities to their personal EEG signal patterns. Based on the currently available sensory hardware, tasks without too much moving or talking could be a good start. A unique role of EEG coach could be helpful in such experiments to encourage more end-users to get involved in such experiments. The short-term goal of these datasets is to inspire new machine learning approaches for decoding behavior from EEG.

4.1 Future Work

Neural interfaces are becoming of increasing interest to industry and having large available datasets could be useful for students and researchers to tease out signals from noisy data. As non-invasive neural recordings become ubiquitous, there is a greater need for such algorithms and datasets. We will continue to focus on developing a framework to make it easier for non-expert end-users to use EEG-based BCI. Our short-term exploration includes developing more specific role sets for the BCI research and development framework, with the emphasis on the role of EEG coach and an online EEG experience community. The impact of continuous feedback to end-users is also a topic we are working on. Also, the idea of step out of the lab to home, starting with encouraging end-users to record EEG during tasks of her/his choice as many times as possible at home, is an interesting direction we are heading to.

4.2 Broader Impact

This approach could contribute to the building of a large-scale EEG dataset using low-cost tools and simple experimental settings at home. Our framework could be illuminated to a broader audience of other time-serious human sensory data collection. After all, brain signals are just one type of sensory health signals, the development of wearable devices are expanding rapidly to provide more perspective about human health information from both real-time monitoring and afterward data analysis.

Together with other human sensory data, EEG-based BCI has the potential to significantly change the ways of human interaction with the rest of the world, including both other individuals, and all the technology devices we developed. The human brain is a type of high-speed neural network, and the current AI-enhanced internet is also a high-speed network, how to connect the two high-speed networks could be an interesting long-term research direction. Our pilot study of quickly gathering large-scale EEG data could be a baby step moving towards this direction.

5 Conclusion

In this paper, we present a framework to gather large-scale EEG data through cheap means and non-expert participants, including experimental design, data collection, data analysis, and community building approaches. Two existing datasets are used as case studies for the framework: Think-Count-Recall (TCR) and Read-Write-Type (RWT). This could be a building block towards the future of everyone using non-invasive, wireless, and affordable BCI systems every day, similar to current smartphone usage for the general non-expert population.

A Appendix

The details of data collection, data analysis, and benchmark machine learning algorithms are in our earlier papers (Qu et al. [2020a,b]),we described the details about how to recruit the fourteen (RWT) or sixteen (TCR) subjects under IRB requirements, the data collection process, data cleaning and feature extraction, and results from benchmark machine learning and deep learning algorithms. We also attached an updated version in the appendix section of this paper to allow readers to replicate these experiments.

A.1 Experiment: Read-Write-Type (RWT)

All subjects first signed an informed consent form. Then, researchers helped them to put on the Muse headbands and test the recording. The Subjects then completed an entrance survey on the computer and became familiar with the online Qualtrics system used in this experiment, especially the sample task switching notice. Next, the Official EEG recording began. A survey in Qualtrics kept track of the time and alerted the subjects to change their tasks after every 60 seconds. After subjects completed all the five tasks, the Official EEG recording stopped and subjects completed a short exit survey.

Subjects: Using experiment TCR as an example, sixteen healthy subjects participated the experiment. Of those, data from three subjects were excluded from subsequent analysis; one for failing to participate in one of the required six sessions, and another because of considerable data loss from one of the Muse electrodes, and the third due to a very high level of noise in the electrode recordings.

Seven males and six females are included in the final data set. Ten of the retained subjects were undergraduate students, the other three were graduate students. Eight subjects were computer science majors. The average age of the subjects was 20.9.

Feature Extraction: We used the absolute Band Powers (BP) feature of the Muse headset, it is the logarithm of the power spectral density of EEG signals summed over that frequency range. Lotte et al. [2018a]. The Muse headsets, are using four dry input electrodes, locations corresponded to sites TP9, AF7, AF8, and T10. The Muse EEG recording application automatically filtered out muscle artifacts, such as eye blinking. Spectral analysis was performed on-board the Muse device and then transmitted at 10 Hz to the EEG recording application on the researcher's computer. Each of these spectral snapshots consists of 20 numeric values – five spectral values for each of the four electrodes.

Data cleaning: During the EEG recording, some electrodes may have temporarily lost contact with the subjects' scalp. The result was that multiple sequential spectral snapshots from one or more electrodes had exactly the same value. When we detected this anomaly, we set that entire spectral snapshot of 20 values to 0, while keeping the time-stamped value, even if the anomaly was only detected on one of the four electrodes. Such data cleaning action resulted in a loss of 27% of the entire data.

Cross Validation: EEG data point samples, if randomly selected, could be near to each other chronologically in both the training set and the testing set. This may cause over-fitting because EEG signals changes slowly. To lessen this possible effect, we first adopted the time-wise cross validation ([Qu et al., 2018b]).

For each five minute session there are five tasks, we divided each tasks to 10 parts, evenly and contiguously, each part has 10% of the data.

Then we did a 10 fold cross validation first and realized that the first 30% of the data were predicted with low accuracy due to a task transition effect. We then cut off these transition times and only used the rest (70%) of the data. In each fold, We trained on six of the remaining seven subsets and tested on the left-out subset. The results reflect some general patterns.

Based on that We also did a session-wise cross validation, to see how the classifiers work with the data from unseen session.

A.2 Experiment: Think-Count-Recall (TCR)

In this experiment, scalp-EEG signals were recorded from sixteen subjects. Each one was tested in six sessions, each session is five minutes long, with five tasks, each task is one minute. Tasks were selected by the subjects together with the researchers, based on frequent tasks in study environments for students in their everyday life. Each subject completed six sessions over several weeks.

Each subject first signed the informed consent form. Then, they put on the Muse headbands and test the recording. Subjects then completed an entrance survey. After these preliminaries, Official EEG recording began. Subjects were directed by an online data collection system, which kept track of time and alerted the subjects to change their tasks after every 60 seconds. After subjects completed all the five tasks, the EEG recording stopped, subjects then completed a short exit survey.

Data cleaning: When collecting EEG data, one or more electrodes may have momentarily lost contact with the subjects' scalp. The result was that multiple sequential spectral snapshots from one or more electrodes had exactly the same 32 bit value. When we detected this anomaly, we set that entire spectral snapshot of 20 values to 0, while keeping the time-stamped value, even if the anomaly was only detected on one electrode. Such cleaning action resulted in a loss of 43% of the entire data. This result echoed with other researches facing the same challenge of low signal-to-noise ratio.

Subjects: Sixteen healthy subjects finished the experiment. Data from four subjects have less than 35 percent data points left after removing noises. Thus these four subjects were excluded from subsequent analysis.

Six males and six females are included in the final data set. Ten of the twelve retained subjects were undergraduate students, the other two were graduate students. Seven subjects were computer science majors; the remaining five were math, biology or psychology majors, or had not yet decided on a field of concentration. The average age of the subjects was 20.2.

All twelve subjects completed the six sessions, producing a data set comprising 360 minutes of EEG recordings (12 subjects x 6 sessions per subject x 5 minutes per session).

Feature Extraction and Feature Selection: We used the Band Powers (BP) features, the absolute band power for a given frequency range (for instance, alpha, 9-13 Hz) is the logarithm of the power spectral density of EEG signals summed over that frequency range. Lotte et al. [2018a]. The Muse headsets are equipped with seven dry electrodes that make contact with the subjects' scalp, three of them are reference, the other four are input. The four input electrode locations corresponded to sites TP9, AF7, AF8, and T10 [Seeck et al., 2017]. The Muse EEG recording application automatically filtered out muscle artifacts, such as eye blinking and jaw movements. The EEG system down-sampled sensor signals from 12k Hz to 220 Hz, with 2uV (RMS) noise. Spectral analysis was performed on-board the Muse device and then transmitted wirelessly at 10 Hz to the researcher's workstation. Each of these spectral snapshots consists of 20 numeric values – five spectral values for each of the four electrodes. This procedure generated a total of 3,000 spectral snapshots per subject per session (10 snapshots/second * 300 seconds).

A.3 Other two experiments

The other two experiments, Python-Math (Qu et al. [2018b]) and GRE-Relax (Qu et al. [2018a]), are using a similar but not so mature approach compare to the newer ones, the details are in our previous papers, and we recommend using the new approach exemplified by experiments TCR and RWT in this paper.

A.4 How to pick thresholds

Detect Noise: Percentage speaking, noise, unknown tasks, and known tasks add up to 100 percent, here we use 1 to represent all the three types together, as shown in Equation 1. During the EEG data collection, one or more electrodes may have momentarily lost contact with the subjects' scalp, especially the TP9 and TP10 electrodes behind ears. The result was that multiple sequential spectral snapshots from one or more electrodes had exactly the same value.

$$noise + unknown + known = 1 \quad (1)$$

We applied Human-In-The-Loop method to determine the noise threshold for how long we should consider such a drop of signals as noise. As shown in Equation 2, we select the time slots which continue noise length are larger than the noise threshold. The total amount of noise is:

$$noise = \sum N(t > nt) \tag{2}$$

Detect Unknown Tasks

After removing the noise through a plateau threshold, we aim to detect the unknown tasks, where unknown tasks refer to those mental activities that might not belong to the five known tasks included in the experimental design.

Here we use experiment TCR as an example, first we implemented unsupervised learning (K-means) to detect the clusters.

We treat each 1/10 second of EEG signal as a 20-dimension data point, and use K-means to find the clusters based on the least squared Euclidean distance. We assume each cluster may represent a certain task, either one of the five known tasks, or a new unknown task not included in the original experimental design. We use subject 1 as an example. In subject one, 3000 data points are recorded in each one of the six five-minute sessions (30 minutes and 18,000 data points for six sessions in total), and the k-means algorithm is looking for clusters in these 18,000 data points. The larger the number of clusters (K), usually the fewer data points in each cluster.

The unknown threshold is defined as the percentage of data points in a certain K-means cluster that represent a known task. For example, when the unknown threshold is 0.5, that means if the number of any one of the five known tasks is more than fifty percent of the total data points, this cluster is considered to be this known task of the highest percentage. If none of the five known tasks reach this 0.5 unknown threshold, we consider this cluster an unknown task because none of the known tasks is dominant in this cluster. Time-wise speaking, that means the data points in this cluster come from different designed known task periods, so they may not belong to any of the known tasks.

Here we can see the prediction accuracy of the known tasks is negatively correlated to the data remain of the known tasks. In other words, with more data points have been recognized as unknown tasks, the prediction accuracy will be higher just using the cleaner version of data points that remain as the known tasks. This pattern is consistent across all of the sixteen subjects. Result with higher accuracy or higher remain can be generated according to demand by using other pairs of K-unknown-threshold combination.

The lower bounds are set as 0.65 for accuracy and 0.34 for data remain. For the accuracy lower bound, accuracy is around 0.65 when only the noise has been removed and no data points have been labeled as unknown tasks, making only accuracy higher than 0.65 has the value to compare. For the data remain lower bound, 0.34 is about one-third of the data points remain, that is to say, less than two-third of data points have been labeled as unknown tasks. Although the prediction accuracy is as high as 86 percent and even higher with data remain less than 0.34, it does not seem to be representative enough for this entire data set.

```
noiseRemoved_data = readcsv("subjID.csv");

for threshold = 0.3:0.02:0.6
    % removed unknown using kmeans
    unknownRemoved_data = Kmeans(noiseRemoved_data,threshold);
    % calulate REMAIN
    REMAIN = size(unknownRemoved_data)/size(original_data);
    % use randomforest to predict tasks
    prediction = RandomForest.fit(unknownRemoved_data);
    % calulate ACCURACY
    ACCURACY = compare(prediction,label);
end

plot2D(ACCURACY,REMAIN,"sort","descend ACCURACY ");
```

Figure 10: Code Example

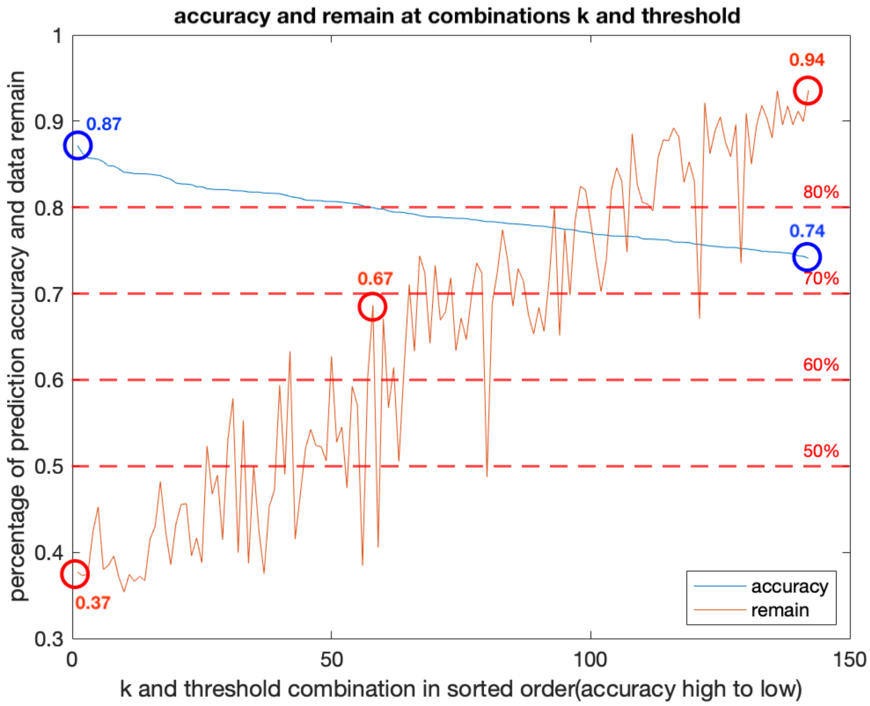

Figure 11: Trade-off of task prediction accuracy and data remain

Trade off between accuracy and data remain

For interpretability, we balance between chasing for higher accuracy and retain a meaningful amount of data. As Figure 11 shows, for subject one in TCR experiment, we selected thresholds to balance the prediction accuracy and data remain for the known tasks. There are five tasks in this TCR experiment example, so the random is 20 percent. The task prediction accuracy can reach 87 percent if the data remain is just 37 percent. While the accuracy remains 74 percent when the data remain is 94 percent.

Figure 10 is the related code to generate Figure 11, X-axis is ordered by task prediction accuracy decreased, thus we can see the trade-off trend. The run time for this step is several seconds on a non-special personal computer. Then the EEG coach and end-users can brainstorm ways to minimize the noise and unknown tasks in future sessions.

References

J. J. Bird, L. J. Manso, E. P. Ribeiro, A. Ekart, and D. R. Faria. A study on mental state classification using eeg-based brain-machine interface. In *2018 International Conference on Intelligent Systems (IS)*, pages 795–800. IEEE, 2018.

J.-A. Chevalier, A. Gramfort, J. Salmon, and B. Thirion. Statistical control for spatio-temporal meg/eeg source imaging with desparsified multi-task lasso. *arXiv preprint arXiv:2009.14310*, 2020.

D. Coyle, J. Principe, F. Lotte, and A. Nijholt. Guest editorial: Brain/neuronal-computer game interfaces and interaction. *IEEE Transactions on Computational Intelligence and AI in games*, 5 (2):77–81, 2013.

A. Craik, Y. He, and J. L. Contreras-Vidal. Deep learning for electroencephalogram (eeg) classification tasks: a review. *Journal of neural engineering*, 16(3):031001, 2019.

D. Devlaminck, W. Waegeman, B. Bauwens, B. Wyns, P. Santens, and G. Otte. From circular ordinal regression to multilabel classification. In *Proceedings of the 2010 Workshop on Preference Learning (European Conference on Machine Learning, ECML)*, page 15, 2010.

L. J. Gabard-Durnam, A. S. Mendez Leal, C. L. Wilkinson, and A. R. Levin. The harvard automated processing pipeline for electroencephalography (happe): standardized processing software for developmental and high-artifact data. *Frontiers in neuroscience*, 12:97, 2018.

M. Ienca, P. Haselager, and E. J. Emanuel. Brain leaks and consumer neurotechnology. *Nature biotechnology*, 36(9):805–810, 2018.

M. J. Kahana, E. V. Aggarwal, and T. D. Phan. The variability puzzle in human memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 44(12):1857, 2018.

M. Kaya, M. K. Binli, E. Ozbay, H. Yanar, and Y. Mishchenko. A large electroencephalographic motor imagery dataset for electroencephalographic brain computer interfaces. *Scientific data*, 5(1): 1–16, 2018.

A. Kübler, E. M. Holz, A. Riccio, C. Zickler, T. Kaufmann, S. C. Kleih, P. Staiger-Sälzer, L. Desideri, E.-J. Hoogerwerf, and D. Mattia. The user-centered design as novel perspective for evaluating the usability of bci-controlled applications. *PLoS One*, 9(12):e112392, 2014.

J. LaRocco, M. D. Le, and D.-G. Paeng. A systemic review of available low-cost eeg headsets used for drowsiness detection. *Frontiers in neuroinformatics*, 14, 2020.

F. Lotte. Signal processing approaches to minimize or suppress calibration time in oscillatory activity-based brain–computer interfaces. *Proceedings of the IEEE*, 103(6):871–890, 2015.

F. Lotte, M. Congedo, A. Lécuyer, F. Lamarche, and B. Arnaldi. A review of classification algorithms for EEG-based brain–computer interfaces. *Journal of neural engineering*, 4(2):R1, 2007.

F. Lotte, L. Bougrain, A. Cichocki, M. Clerc, M. Congedo, A. Rakotomamonjy, and F. Yger. A review of classification algorithms for EEG-based brain–computer interfaces: a 10 year update. *Journal of neural engineering*, 15(3):031005, 2018a.

F. Lotte, C. Jeunet, J. Mladenović, B. N'Kaoua, and L. Pillette. A bci challenge for the signal processing community: considering the user in the loop, 2018b.

K. J. Miller. A library of human electrocorticographic data and analyses. *Nature human behaviour*, 3 (11):1225–1235, 2019.

T. E. Nichols, S. Das, S. B. Eickhoff, A. C. Evans, T. Glatard, M. Hanke, N. Kriegeskorte, M. P. Milham, R. A. Poldrack, J.-B. Poline, et al. Best practices in data analysis and sharing in neuroimaging using mri. *Nature neuroscience*, 20(3):299–303, 2017.

C. R. Pernet, S. Appelhoff, K. J. Gorgolewski, G. Flandin, C. Phillips, A. Delorme, and R. Oostenveld. Eeg-bids, an extension to the brain imaging data structure for electroencephalography. *Scientific data*, 6(1):1–5, 2019.

X. Qu, M. Hall, Y. Sun, R. Sekuler, and T. J. Hickey. A personalized reading coach using wearable EEG sensors-a pilot study of brainwave learning analytics. In *CSEDU (2)*, pages 501–507, 2018a.

X. Qu, Y. Sun, R. Sekuler, and T. Hickey. EEG markers of stem learning. In *2018 IEEE Frontiers in Education Conference (FIE)*, pages 1–9. IEEE, 2018b.

X. Qu, P. Liu, Z. Li, and T. Hickey. Multi-class time continuity voting for eeg classification. In *International Conference on Brain Function Assessment in Learning*, pages 24–33. Springer, 2020a.

X. Qu, Q. Mei, P. Liu, and T. Hickey. Using eeg to distinguish between writing and typing for the same cognitive task. In *International Conference on Brain Function Assessment in Learning*, pages 66–74. Springer, 2020b.

A. Roc, L. Pillette, J. Mladenovic, C. Benaroch, B. N'Kaoua, C. Jeunet, and F. Lotte. A review of user training methods in brain computer interfaces based on mental tasks. *Journal of Neural Engineering*, 2020.

Y. Roy, H. Banville, I. Albuquerque, A. Gramfort, T. H. Falk, and J. Faubert. Deep learning-based electroencephalography analysis: a systematic review. *Journal of neural engineering*, 16(5): 051001, 2019.

D. Sabbagh, P. Ablin, G. Varoquaux, A. Gramfort, and D. A. Engemann. Manifold-regression to predict from meg/eeg brain signals without source modeling. *arXiv preprint arXiv:1906.02687*, 2019.

G. Schalk, D. J. McFarland, T. Hinterberger, N. Birbaumer, and J. R. Wolpaw. Bci2000: a general-purpose brain-computer interface (bci) system. *IEEE Transactions on biomedical engineering*, 51 (6):1034–1043, 2004.

M. Seeck, L. Koessler, T. Bast, F. Leijten, C. Michel, C. Baumgartner, B. He, and S. Beniczky. The standardized EEG electrode array of the IFCN. *Clinical neurophysiology*, 128(10):2070–2077, 2017.

K. Sierra and B. Bates. *Head first java*. " O'Reilly Media, Inc.", 2003.

Y. R. Tabar and U. Halici. A novel deep learning approach for classification of eeg motor imagery signals. *Journal of neural engineering*, 14(1):016003, 2016.

T. Tu, J. Paisley, S. Haufe, and P. Sajda. A state-space model for inferring effective connectivity of latent neural dynamics from simultaneous eeg/fmri. *Advances in Neural Information Processing Systems*, 32:4662–4671, 2019.

X. Zhang, L. Yao, X. Wang, J. J. Monaghan, D. Mcalpine, and Y. Zhang. A survey on deep learning-based non-invasive brain signals: recent advances and new frontiers. *Journal of Neural Engineering*, 2020.