Automated Classification of EEG Signals for Predicting Students' Cognitive State during Learning

Xi Liu, Pang-Ning Tan
Michigan State University
{liuxi4,ptan}@cse.msu.edu

Lei Liu*
Huawei Technologies, USA
leiliu@ieee.org

Steven J. Simske
HP Labs, USA
steven.simske@hp.com

ABSTRACT

For distance learning applications, inferring the cognitive states of students, particularly, their concentration and comprehension levels during instruction, is important to assess their learning efficacy. In this paper, we investigated the feasibility of using EEG recordings generated from an off-the-shelf, wearable device to automatically classify the cognitive states of students as they were asked to perform a series of reading and question answering tasks. We showed that the EEG data can effectively predict whether a student is attentive or distracted as well as the student's reading speed, which is an important measure of reading fluency. However, the EEG signals alone are insufficient to predict how well the students can correctly answer questions related to the reading materials as there were other confounding factors, such as the students' background knowledge, that must be taken into consideration. We also showed that the accuracy in predicting the different cognitive states depends on the choice of classifier used (global, local, or multi-task learning). For example, the concentration level of a student can be accurately predicted using a local model whereas a global model that incorporates side information about the student's background knowledge is more effective at predicting whether the student will correctly answer questions about the materials they read.

CCS CONCEPTS

• **Information systems → Data mining**;

ACM Reference format:
Xi Liu, Pang-Ning Tan, Lei Liu, and Steven J. Simske. 2017. Automated Classification of EEG Signals for Predicting Students' Cognitive State during Learning. In *Proceedings of WI '17, Leipzig, Germany, August 23-26, 2017,* 8 pages.
https://doi.org/10.1145/3106426.3106453

1 INTRODUCTION

The use of technology in education has become increasingly prevalent to aid instructors in making their lectures more vivid and the learning materials more accessible to students. However, the use of technology to support real-time monitoring of how well students grasp the materials they learn is still in its infancy. The ability to measure the student's concentration and comprehension levels during instruction is essential as they provide valuable feedback to the instructor whether the learning goals have been achieved and to identify the type of intervention needed to improve learning outcomes. This has led to the growing interest in developing Brain Computer Interface (BCI) systems to collect and analyze signals from the human brain to determine its cognitive state [10] [16]. The advantage of using brain signals is that they provide an objective way to measure the students' current state of mind compared to the traditional approach where instructors have to subjectively gauge the students' concentration and understanding based on their facial and bodily expressions.

In recent years, there has been increasing interest in utilizing EEG signals to determine the cognitive state of students as they engaged in various learning activities. Traditional EEG devices are expensive and bulky, making it cumbersome to conduct experimental studies in a non-controlled environment. However, with the rapid advances in sensing technology, new generation of EEG devices that are non-invasive, portable, and affordable have been developed, thus providing an opportunity to conduct measurement studies where the human subjects are immersed in a real-world environment. In this paper, we investigate the feasibility of using an off-the-shelf, wearable EEG device to monitor the concentration and understanding levels of human subjects as they were asked to perform a series of reading and question answering tasks.

Previous studies using EEG signals for classifying cognitive state such as attention have focused mostly on detecting changes in the brain signals for a limited subset of the band waves [7][4][16]. They have also considered using only a limited number of features extracted from the EEG recordings (e.g., root mean square of energy value for a particular band wave). More importantly, none of the previous studies have considered predicting the comprehension level using EEG data. Other issues such as the choice of classifier used and how it affects classification performance have not been sufficiently investigated in the past. To address these concerns, we have extracted a broad set of features from various frequency bands and assess their relevance to the classification task. We have also compared the relative performance of local, global, and multi-task learning models to determine their effectiveness at discriminating the different cognitive states. A multi-task learning approach simultaneously learns the local model for each human subject by solving a joint optimization problem, instead of learning them independently. It can thus be regarded as a hybrid between the local and global models.

We conducted a study in which each human subject was asked to perform a series of reading and question answering tasks. The

*This research was conducted while Lei Liu was at HP Labs.

WI '17, August 23-26, 2017, Leipzig, Germany

ACM ISBN 978-1-4503-4951-2/17/08…$15.00
https://doi.org/10.1145/3106426.3106453

Table 1: Summary of EEG classification tasks.

Cognitive state	Binary classes	Mental task
Concentration level	Attentive/ Distracted	Reading task
Reading speed	Fast/slow	Reading task
Recall level	Good/Poor	Question-answering task

EEG data collected from this study were used to develop predictive models for three cognitive state classification tasks summarized in Table 1. The classification tasks are useful as they provide valuable feedback that can be used to improve learning outcomes. For example, accurate classification of concentration level can help the instructors to determine whether the students paid careful attention to the materials that were taught. Second, reading speed is an important indicator of reading fluency or the decoding skills of readers [8]. Reading speed is measured as a ratio between the duration in which a subject spent on the reading material and the number of words in the given passage. A threshold is set to decide whether the reading speed is high or low. For example, if a student spends, on average, 200 words per minute to read a given passage, then a fast reading speed, say, above 250 words per minute, may indicate that the passage can be easily comprehended by the student or that the student tries to quickly skim through the material without focusing on its details. Finally, the ability to correctly answer questions related to the reading materials provide a good indication about the reading comprehension skills of the student.

A major challenge to this research is obtaining reliable ground truth labels for the various classification tasks. For example, it is hard to tell whether a subject is attentive or distracted during a reading session. In [11], videos of the subjects' facial expressions were recorded for experts to interpret whether they were paying attention. Such visual inspection is very subjective and is prone to misinterpretation. To overcome this problem, following the approach used in [10], the EEG measurements were taken both during reading and non-reading ("mind-wandering") sessions to distinguish between attentive and distracted states of mind.

In summary, the main contributions of this work are as follows:

(1) We investigated the feasibility of using EEG data to infer various cognitive states of human subjects. Our results suggested that EEG data can reliably predict the concentration and reading speed levels of the subjects. However, predicting recall level is more challenging as it can be affected by other external factors (e.g., proficiency and background of the subjects).
(2) We designed experiments to collect ground truth labels for the various cognitive state classification tasks.
(3) We compared the performance of various types of classifiers (local, global, and multi-task learning) in terms of their accuracy in distinguishing different cognitive states. We showed that a local model is sufficient to predict the students' concentration level. However, a global model that incorporates side information is more effective at predicting recall level (though its accuracy is lower than those for concentration and reading speed level predictions).

2 RELATED WORK

This section reviews some of the previous works on the application of EEG data for educational purposes. Recent progress in neuroscience has given us a deeper understanding of how the brain works and provides us with novel ways to detect and analyze brain activities. While there are other brain activity monitoring and imaging techniques available beside EEG, many of them, such as MEG (magnetoencepha-lography) and fMRI, require strict working conditions (e.g. in a shielded room), special equipments (such as liquid helium-cooled detectors), and licensed experts to operate the machine. EEG devices are easier to use and non-invasive compared to others such as electrocorticography, which make them more suitable to be deployed in a classroom or other real-world environment. For this reason, EEG has been widely used, not only in a clinical setting, but also to aid in neuroscience, cognitive science, and psychology research. Various EEG studies have been conducted to explain human emotions and expressions [14] [4], as well as measuring the cognitive load when performing tasks such as driving [15] and studying [3].

Monitoring the cognitive states of students is particularly useful for applications such as distance learning due to the lack of face-to-face interactions between the instructor and students. Previous research has shown that cognitive efforts such as concentration was found to be highly correlated with EEG frequencies [16] [5] [11]. Nevertheless, obtaining the ground truth label about the concentration level of a human subject is harder to determine. In [11], the participants were subjected to two different scenarios: learning with and without distractions, in which the former was used to represent concentrated state of mind while the latter represents distracted state. In [9], the participants were asked to rate their own perceived level of concentration with SAM (Self-Assessment Manikin) test [13]. The self-assessment is used to provide the ground label of the data. However, labels obtained based on self reporting is harder to use since they vary from one subject to another. In [11], participants' facial expressions were also recorded for experts to determine their concentration level. This approach is highly subjective and prone to misinterpretation.

Previous studies are also limited in that they focus on EEG signals from a limited number of band waves to determine the cognitive state of the subjects. In [7], only the α band wave was investigated for cognitive loads whereas in [9], the α and θ band waves were extracted as features for concentration classification. There is also a commonly used index of concentration level known as E-signals [16], which is based on the α, β and θ band waves. None of these studies consider the complete spectrum of the EEG signals. Many of the previous studies also use the raw values of the band waves to train their models. For example, in [16], the task engagement index was calculated using the magnitude of the band waves, while in [14], the signals from each channel electrode was directly used as features. Finally, none of the previous studies consider using EEG data to monitor the understanding level of students.

3 METHODOLOGY

This section describes our methodology for using EEG data to infer the cognitive states of human subjects. The EEG data was obtained using an adjustable, lightweight EEG headset called Muse [1]. The

Table 2: Basic Information of the Participants.

Subject ID	Gender	English Proficiency	Technical Background
1	Male	High (native)	Mid
2	Male	High	High
3	Male	Low	High
4	Male	High	High
5	Female	High (native)	Low
6	Male	High	High
7	Male	Mid	High
8	Male	Low	Low
9	Female	Mid	Low
10	Female	Low	Low
11	Female	Low	Mid

Muse headset has 7 sensors capable of reading 4 channels of data—two on the forehead (Fp1 and Fp2) and two behind the ears (TP9 and TP10). The EEG signals were generated at a sampling rate of 220Hz. Muse not only collects the raw EEG signals, it also has a digital signal processing component for de-noising and decomposing the time series into measurements of power spectral density.

3.1 Design of Experiment

The goal of our study is to investigate the feasibility of using EEG recordings to determine the cognitive state of the human subjects as they were provided with some passages to read. Each subject was initially invited into a quiet room. An instructor was present to help the subject put on the Muse device properly and to pair the Muse device with a computer via Bluetooth. The MuseLab software was installed on the computer to visualize and record the EEG signals. Each subject was required to perform the following tasks:

- Reading: The subject was given a passage to read from the abstract of an article that appeared in the *Science* Journal. The articles belong to many subject areas including environmental science, education, physics, and cell biology. The abstracts chosen do not require extensive technical background to understand their content. Each subject was allocated two minutes to complete the reading task.
- Question answering: Each reading task was immediately followed by a question-answering task. The subjects were required to answer the following 3 multiple-choice questions related to the reading material:
  (1) Which subject area best describes this article?
  (2) What is this article mainly about?
  (3) Which of the following is NOT mentioned in this article?
  The 3 questions were ordered in increasing level of difficulty to assess the recall level of the subjects, i.e., whether they understood the passage they had read.
- Mind wandering: With their eyes open, the subjects were asked to stay in a comfortable pose. They were allowed to look around, but were asked not to focus their thought on anything for a duration of one minute. This mind-wondering session is used to capture EEG signals when the subject is not concentrating.

Table 3: Schedule of mental task experiment.

ID	Task	ID	Task
A1-1, R1-1	Reading, Q&A	A2-1, R2-1	Reading, Q&A
A1-2, R1-2	Reading, Q&A	A2-2, R2-2	Reading, Q&A
M1-3	Mind wandering	M2-3	Mind wandering
A1-4, R1-4	Reading, Q&A	A2-4, R2-4	Reading, Q&A
A1-5, R1-5	Reading, Q&A	A2-5, R2-5	Reading, Q&A
M1-6	Mind wandering	M2-6	Mind wandering
A1-7, R1-7	Reading, Q&A	A2-7, R2-7	Reading, Q&A
A1-8, R1-8	Reading, Q&A	A2-8, R2-8	Reading, Q&A
M1-9	Mind wandering	M2-9	Mind wandering
A1-10, R1-10	Reading, Q&A	A2-10, R2-10	Reading, Q&A
A1-11, R1-11	Reading, Q&A	A2-11, R2-11	Reading, Q&A
M1-12	Mind wandering	M2-12	Mind wandering
A1-13, R1-13	Reading, Q&A	A2-13, R2-13	Reading, Q&A
A1-14, R1-14	Reading, Q&A	A2-14, R2-14	Reading, Q&A
M1-15	Mind wandering	M2-15	Mind wandering
A1-16, R1-16	Reading, Q&A	A2-16, R2-16	Reading, Q&A
A1-17, R1-17	Reading, Q&A	A2-17, R2-17	Reading, Q&A
M1-18	Mind wandering	M2-18	Mind wandering
A1-19, R1-19	Reading, Q&A	A2-19, R2-19	Reading, Q&A
A1-20, R1-20	Reading, Q&A	A2-20, R2-20	Reading, Q&A
M1-21	Mind wandering	M2-21	Mind wandering
A1-22, R1-22	Reading, Q&A	A2-22, R2-22	Reading, Q&A
M1-23	Mind wandering	M2-23	Mind wandering

There were 11 volunteer subjects who participated in the study. All subjects were healthy, right-handed, with varying degrees of English proficiency and technical background, as shown in Table 2. Each subject was required to complete 30 reading/question-answering and 16 mind wandering trials. We divided the trials into 2 separate sessions (conducted on two different days) to avoid over-taxing the subjects. Table 3 presents the complete schedule of our mental task experiment.

3.2 Data Collection

We collected EEG recordings from each subject for all 76 mental tasks (30 reading, 30 question answering, and 16 mind wandering tasks). Previous studies have suggested there is significant relationship between a cognitive task such as concentration and the different frequency bands of EEG signals [7] [12] [6]. Instead of using the raw EEG signals, we use the preprocessed signals extracted by the digital signal processing (DSP) component of Muse. The typical frequency bands used in previous studies are δ (1-4 Hz), θ (5-8 Hz), α (9-13 Hz), β (12-30 Hz) and γ (30-50 Hz). Description about the time series can be found in [1].

As previously noted, Muse generates 4 channels of data, where each channel has its own α, β, γ, δ, and θ time series. We took the average value of the four channels to calculate the average relative power for each frequency band. In addition, we also analyzed the *is_good* signal associated with each channel. Since the data from one or more channels is not always reliable due to various reasons (e.g., the headgear not mounted properly or the Bluetooth is disconnected), the *is_good* time series indicates which channels are

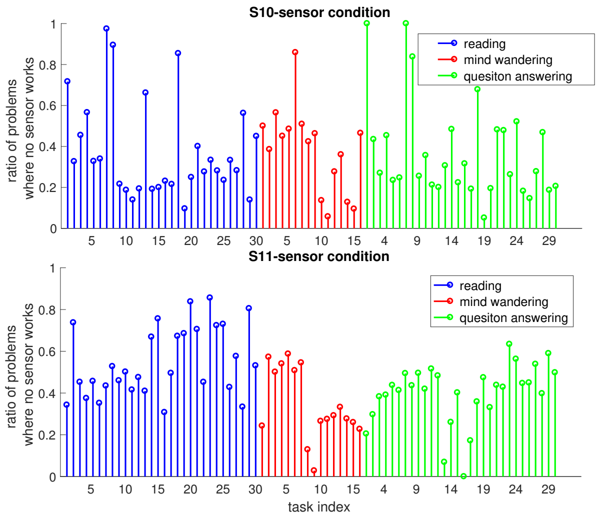

Figure 1: Time series that shows the working condition of Muse for two subjects, with *ID=10* and *ID=11*. The vertical axis represents the percentage of time during which none of the 4 channels are working. The blue stems correspond to reading trials, the green ones are question answering trials, while the red ones are mind wandering trials.

working properly [1]. If the data for a given channel is unreliable, we ignore its value when computing the average relative band power. At least one of the channels must work properly for us to trust the EEG data. Furthermore, as the number of channels that worked properly may vary from time to time, we calculated the percent of time within each trial for which all four channels were not working properly. If the percent of time exceeds 50% for a given trial, we discarded its corresponding EEG data. Figure 1 shows the percentage of time in which the sensors are not operating correctly for two subjects, with IDs #*10* and #*11*, respectively. As most of the data are unreliable, we decided to remove these two subjects from our experimental study. We ended up using EEG data from the remaining 9 subjects in our study.

3.3 Feature Extraction

Figure 2 shows examples of the average α, β, γ, δ, and θ time series from a same subject during one of the reading and mind wandering trials. The EEG signals for the two tasks are quite different, suggesting the possibility of using a classifier to discriminate the two cognitive states. Instead of using the raw time series for classification, we derive the following summary statistics from each of the 5 relative band powers—root-mean-square value, maximum and minimum amplitude, peak-to-peak of amplitude, variance, and approximate entropy—to create 30 features (5 bands $\times$ 6 summary statistics) characterizing the EEG signals.

In addition to the EEG signals, we have also collected side information about the subjects and the reading materials.

- *English proficiency and technical level of the subjects*: Fluency in English may affect the performance of a subject during the reading and question answering sessions. Since the reading materials were obtained from the Science journal, the technical background of the human subjects is expected to have an impact on their understanding level.
- *Reading time and number of words in the abstract*: The reading speed is computed based on the ratio between number of words in a given abstract to reading time.
- *Difficulty level of the reading material*: We apply the Linsear write index [2], a widely-used readability metric for English text. The metric is calculated based on sentence length and the number of words used that have three or more syllables.
- *Difficulty level of the questions*: The multiple choice questions may have different levels of difficulty. To compute the difficulty level of a question for a given human subject, we took the average number of correct answers provided by other human subjects on the same question. If there are few participants who can answer the question correctly, the difficulty level of the question will be high.

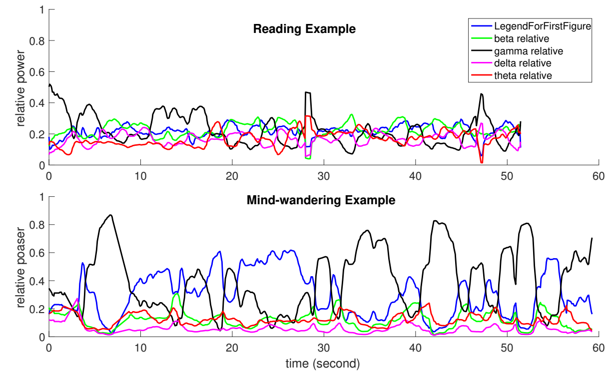

Figure 2: An example of the relative band waves during a reading trial (upper) and a mind-wandering trial (lower).

The features used for the different classification tasks are summarized in Table 4. We use the EEG signals to classify the concentration level and reading speed of the subjects during their reading trials. To classify the recall level, we also use side information about the subject and reading material in addition to the output of the concentration level classifier.

3.4 Classification

We applied the following classifiers to the EEG data:

- Local model. We train a local model for each subject independently using the subject's training data. The test data for each subject is then classified using the local model trained for the subject.
- Global model. We train a global model by pooling together the training data from all the subjects. In this setting, the same model is used to classify the test data for all the subjects.

Table 4: Features used to classify the cognitive states of the human subjects.

Feature category	Features	Concentration Level	Reading Speed	Recall Level
EEG	EEG signals	x	x	
Reading material	Size of article			x
	Linsear index			
Learning process	Concentration			x
	Duration			
	Reading speed			
Subjects' background	English level			x
	Technical level			

- Multi-task learning: In this setting, the test data for each subject is classified using his/her local model. Here, the local models of the subjects are trained by optimizing a joint objective function, as described in Section 3.4.2.

3.4.1 Logistic Regression. We employ the $L1$-regularized logistic regression to train both our local and global classification models. Logistic regression is a widely used linear model for binary classification. The $L1$-regularizer was used to generate sparse models and prevent overfitting. The objective function is given below:

$$\min_{\theta} \sum_{j=1}^{n} \log\left(1 + exp(-y_j(\theta^T \mathbf{x}_j))\right) + \gamma ||\theta||_1 \qquad (1)$$

where θ is the model parameter, γ is the regularizer, n is the total number of training instances, X is the input data matrix, and $\mathbf{y}$ is the class labels (-1 or $+1$).

3.4.2 Multi-task Learning. Multi-task learning is an emerging learning paradigm for solving multiple, related learning tasks jointly by exploiting the common structure of the problem. The approach assumes there is a set of correlated prediction tasks to be solved, where each task can have its own training data. In this study, the prediction of cognitive state for each subject can be regarded as a single learning task. Despite the differences in the EEG signals and other features for each subject, we investigate the feasibility of applying multi-task learning to jointly train their predictive models. Specifically, we use the sparse graph-regularized logistic regression classifier provided by the MALSAR [18] software package. The classifier was designed to optimize the following objective function:

$$\begin{aligned} \min_{\mathbf{\Theta}} \quad & \sum_{i=1}^{t} \sum_{j=1}^{n_i} \log(1 + exp(-\mathbf{Y}_{ij}(\mathbf{\Theta}_i^T \mathbf{X}_{ij}))) + \rho_1 ||\mathbf{\Theta R}||_F^2 + \rho_2 ||\mathbf{\Theta}||_1 \qquad (2) \\ s.t. \quad & \mathbf{R} = \mathbf{I}_t - \mathbf{1}_t / t \end{aligned}$$

where, t is the number of tasks (subjects), $\mathbf{X}_i$ is the data matrix for task i, $\mathbf{Y}_i$ is ground truth labels for task i, $\mathbf{\Theta}$ is the parameter matrix for all t tasks, $\mathbf{\Theta}_i$ is the i-th column of $\mathbf{\Theta}$, or the parameter vector of the i-th task, $\mathbf{I}_t$ is a t-dimensional identity matrix, and $\mathbf{1}_t$ is a t-dimensional square matrix with all elements as 1. The first regularization term, $||\mathbf{\Theta R}||_F^2$, penalizes the deviation of each local model from the mean model $\sum_{i=1}^{t} \theta_i$, while the second regularization term, $||\mathbf{\Theta}||_1$, controls the model sparsity.

4 EXPERIMENTAL EVALUATION

We have conducted extensive experiments to provide answers to the following research questions: (1) Can EEG data be effectively used to recognize different cognitive states? (2) Should we use features from all frequency bands or a subset of the bands? (3) Does the choice of classifier affect prediction accuracy and if so, which choice is most effective for each cognitive state prediction task?

4.1 Ground Truth Labeling

One of the key challenges in this research is to determine the ground truth label for each trial. Our procedure for annotating the EEG data from different trials of each human subject is described below.

Labels for Concentration Level. Following the approach used in [5], we assign all the reading tasks to the positive (attentive) class and the mind wandering tasks to the negative (distracted) class. The proportions of positive and negative examples in the labeled data are 65% and 35%, respectively.

Labels for Reading speed. Previous studies [8, 17] have suggested a strong correlation between reading speed or fluency and reading comprehension. We measure the reading speed for all the abstracts read by each subject. For each subject, if the reading speed is higher than the median, we label the reading trial as positive ("high speed"); otherwise, it is labeled as negative ("low speed") class. There are almost equal proportions of positive and negative examples in the labeled data.

Labels for Recall Level. For predicting recall level, we calculate the number of correct answers provided by the subjects during the question-answering trials. Since most subjects were able to answer the first question correctly, we determine the recall level based on the answers to the last two questions. If the answers to the last two questions were correct, we label the question-answering trial as positive ("good") class; otherwise it is labeled as negative ("poor") class. The proportions of positive and negative examples in the labeled data are 55% and 45%, respectively.

4.2 Experimental Setup

We use repeated k-fold cross validation to report the average classification performance for each method. Specifically, we divide the dataset into k folds and use a subset of the folds for training and the remaining for testing. This process is repeated 10 times. We vary the training set size by setting k to 5 and 10. For example, to obtain 10% training set, we use 10-fold cross validation, with 1 fold reserved for training and the remaining 9 folds for testing. Similarly, for 80% training set, we apply 5-fold cross validation, with 4 folds reserved for training and the remaining fold for testing. For local and multi-task learning classifier, the models were developed using the EEG data associated with each subject. For global classifier, we aggregated the training data from all subjects before constructing the model.

4.3 Performance Evaluation

We reported the performance for the various cognitive state prediction tasks in terms of their model accuracy as well as their F1 scores for both positive and negative classes. Although there were initially 11 human subjects, only 9 of them produce reliable data (see Section

Table 5: Performance comparison for concentration level prediction

		Accuracy (%)	+ class F1 score (%)	- class F1 score (%)
80% data for training	Local	91.88 ± 0.47	93.93 ± 0.36	87.73 ± 0.72
	Global	81.42 ± 0.84	86.20 ± 0.65	71.56 ± 1.39
	Multi-task	91.47 ± 0.48	93.60 ± 0.35	87.22 ± 0.77
20% data for training	Local	85.40 ± 1.07	89.25 ± 0.77	77.27 ± 1.88
	Global	79.47 ± 0.63	84.85 ± 0.46	68.15 ± 1.11
	Multi-task	86.42 ± 1.02	90.00 ± 0.69	78.85 ± 1.87
10% data for training	Local	79.72 ± 0.66	85.34 ± 0.56	67.08 ± 0.88
	Global	76.46 ± 0.78	82.84 ± 0.49	65.52 ± 1.72
	Multi-task	81.33 ± 0.63	86.43 ± 0.48	70.05 ± 1.1
Random Guess	Local	55.01	65.80	34.29
	Global	54.97	65.77	34.23

3.2) that will be used to report our experimental results. We also reported the results for random guessing, which is computed based on the proportion of positive and negative examples in the data. For example, suppose there are n_+ positive and n_- negative examples in the data, the accuracy for random guessing is $(n_+^2 + n_-^2)/(n_+ + n_-)^2$.

4.3.1 Concentration Level Prediction. Table 5 shows the results for concentration level prediction using all three classification methods (local, global, and multi-task learning). With 80% training data, the accuracies for all three methods are high, exceeding 81%, which suggest that concentration level can be accurately determined from the EEG data. Furthermore, both the local and multi-task learning models appear to outperform the global model, achieving an accuracy exceeding 90%. This suggests that the EEG signals that characterize the attentive and distracted classes for each subject is quite different, thus aggregating the data together to build a global model will degrade the overall prediction accuracy.

By decreasing the training set size, the results given in Table 5 also suggest that: 1). the accuracies for all three methods will be lower, which is not surprising; 2). The accuracy for local classifier decreases most rapidly, from 91.88% to 79.72%, for the training set size of 10% since there are only 3-4 training examples available for each local classifier; 3). The accuracy for global model does not decrease as rapidly, from 81.42% to 76.46%, because the model still has adequate training examples (even when only 10% of the data is reserved for training, there are still 30-40 training examples available); 3). The local models still outperform the global model, which again indicates the importance of training personalized models to classify concentration level; 4). Multi-task learning can exploit information from other subjects to produce higher accuracy models compared to the local classifiers when the training set size is small. These results suggest that multi-task learning can produce better models compared to the local and global classifiers when there are limited EEG data available for each subject.

Figure 3 shows the prediction accuracy for each individual subject with 80% training data. Observe that the accuracy for both local and multi-task learning models are higher than the global model, which is consistent with the results shown in Table 5. Furthermore, there is a significant difference in performance between the local and global models for subject IDs #1 and #2. This suggests that

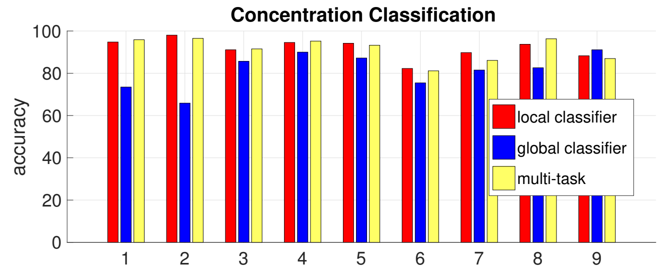

Figure 3: Concentration level prediction accuracy for individual subjects with 80% training data.

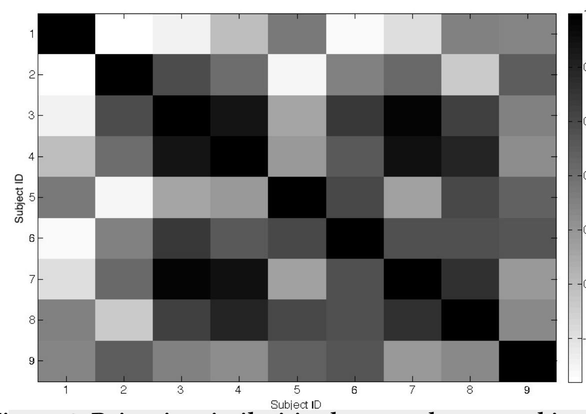

Figure 4: Pair-wise similarities between human subjects in terms of their EEG derived features.

the EEG-derived features characterizing the classes for the two subjects are different from those for other subjects, which is why a global model is not as effective when applied to the EEG data. The finding can be confirmed by looking at the heat map shown in Figure 4, which shows the pairwise similarities between the different subjects. The two subjects IDs #1 and #2 appear to have lower similarities compared to other subjects.

Figure 5 compares the results of using our proposed features against Muse's own predictions [1], which are based on the σ band wave only, and the E-signal features proposed in [16], which were based on the α, β, and θ band waves. The results suggested that our proposed features are more effective than both Muse and E-features irrespective of the type of classifier used. This result suggests the need to use a wider range of frequency spectrum to improve concentration level prediction, unlike other prior studies that are limited to using one or two frequency bands [7] [9].

4.3.2 Reading Speed Prediction. Next, we examine the effectiveness of using EEG data to predict the reading speed of the human subjects. With 80% training set size, the results shown in Table 6 suggest that the EEG features are still quite effective, i.e., significantly better than random guessing, though the accuracies are much lower (around 67%) compared to concentration level prediction (around 90%). Furthermore, all three classifiers (local, global, and multi-task learning) are quite comparable in performance.

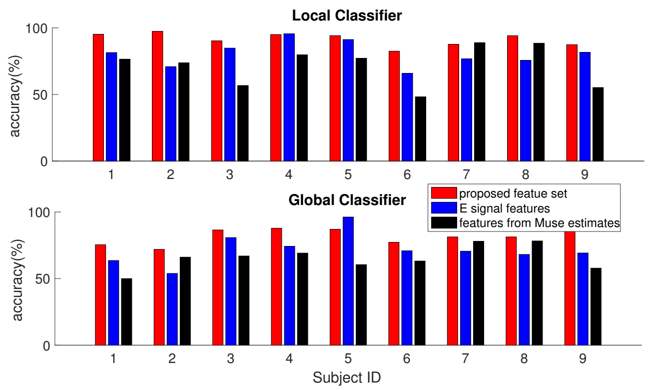

Figure 5: Performance comparison for concentration level prediction using different feature sets.

Table 6: Performance comparison for reading speed prediction.

		Accuracy (%)	+ class F1 score (%)	- class F1 score (%)
80% data for training	Local	68.07 ± 1.69	67.36 ± 1.86	68.74 ± 1.79
	Global	67.92 ± 1.59	68.10 ± 1.76	67.73 ± 1.5
	Multi-task	68.33 ± 1.12	68.16 ± 1.16	68.50 ± 1.2
20% data for training	Local	59.25 ± 1.88	57.64 ± 1.96	60.74 ± 1.86
	Global	63.94 ± 1.1	63.98 ± 1.31	63.87 ± 1.38
	Multi-task	64.37 ± 1.3	64.31 ± 1.25	64.39 ± 1.75
10% data for training	Local	56.75 ± 0.85	56.05 ± 1.33	57.41 ± 0.7
	Global	61.46 ± 1.18	60.78 ± 1.71	62.09 ± 1.26
	Multi-task	61.05 ± 0.85	60.16 ± 1.61	61.88 ± 0.44
Random Guess	Local	50.01	50.39	49.63
	Global	50.00	50.38	49.62

When the training set size decreases, the performances for all three methods will degrade, with the accuracy for the local classifier decreases most rapidly compared to the global classifier and multi-task learning. Unlike concentration level prediction, both the global and multi-task learning models become significantly better than the local models. This is because reading speed classification is a much harder problem, thus, having sufficient training data is important to distinguish fast from slow reading speed.

In addition, Figure 6 shows the prediction results for the individual subjects with 80% training set size. In general, the accuracies are quite comparable, though, for some subjects, the local model is better whereas for others, the global model has a slight advantage.

4.3.3 Recall Level Prediction. Finally, we report the results for predicting recall level of the human subjects. First, we examine the results when the classifier is trained using the EEG-derived features alone. The results shown in Table 7 suggest that the classification performance is not much better than random guessing. This suggests that recall level is harder to predict using EEG data compared to concentration level and reading speed predictions.

Instead of using the raw EEG features, we consider using the output of the concentration level prediction as one of the input features for recall level prediction. We also augment our feature set with side

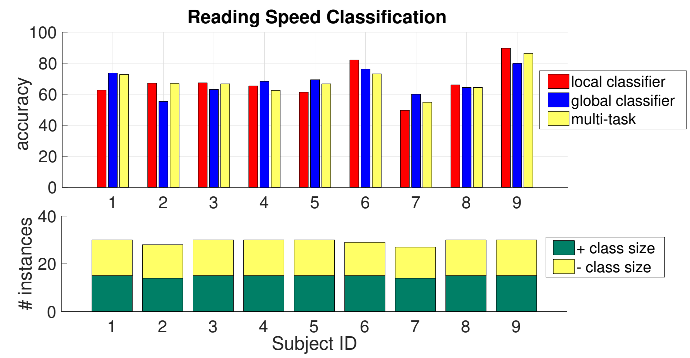

Figure 6: Reading speed prediction results for individual subjects with 80% training set size.

Table 7: Performance of recall level prediction using only EEG signals.

	Accuracy (%)	+ class F1 score (%)	- class F1 score (%)
Local	57.81 ± 2.25	63.27 ± 2.02	50.40 ± 2.82
Global	51.45 ± 2.35	58.28 ± 2.34	41.86 ± 3.17
Multi-task	57.55 ± 1.97	64.02 ± 1.63	48.18 ± 3.15
Random guess (local)	53.36	57.61	48.15
Random guess (global)	50.50	55.02	44.98

information about the reading materials and the human subjects. Table 8 summarizes the classification results. Note that, among all the side information features listed in Table 4, the local and multi-task learning models include only features derived from the reading material and learning process, whereas the global model also incorporates features about the individual subjects' background. The additional features help the global classifier to outperform both the local and multi-task learning models. Note that if we exclude the individual subjects' background, the accuracy for the global model decreases from 63.12% to 59.07%, while the accuracy for the local models remains the same.

Observe that the performance for all three classifiers are better than the results shown in Table 7, which employ only the EEG features. In addition, the global model achieves 63% accuracy, which is significantly higher than random guessing as well as the local models, though it is still lower than the accuracies for concentration level and reading speed predictions. Similar to reading speed classification, the performances for all three methods degrade considerably with decreasing training set size.

The predictive performance for individual subjects shown in Figure 7 validated our results that the global model is generally more effective than the other two competing models. This result makes sense as the global classifier is the only one that can utilize side information about the individual subjects, such as technical background and English proficiency, to predict recall level of the subjects. As shown in Figure 8, these are the two most prominent features for classification using the global model.

Table 8: Performance comparison for recall level prediction using side information.

		Accuracy (%)	+ class F1 score (%)	- class F1 score (%)
80% data for training	Local	61.52 ± 1.67	66.99 ± 1.66	53.90 ± 2.30
	Global	63.12 ± 0.96	69.21 ± 0.86	54.03 ± 1.36
	Multi-task	62.12 ± 1.30	68.09 ± 1.19	53.41 ± 1.63
20% data for training	Local	54.77 ± 1.85	58.80 ± 2.32	49.80 ± 1.76
	Global	59.72 ± 1.11	65.64 ± 1.22	51.30 ± 1.45
	Multi-task	59.29 ± 1.40	63.28 ± 1.59	54.28 ± 1.85
10% data for training	Local	54.30 ± 0.66	59.33 ± 0.98	47.81 ± 1.30
	Global	57.78 ± 1.33	63.00 ± 1.3	50.80 ± 1.90
	Multi-task	55.68 ± 0.94	60.63 ± 0.87	49.24 ± 1.93
Random Guess	Local	53.36	57.61	48.15
	Global	50.50	55.02	44.98

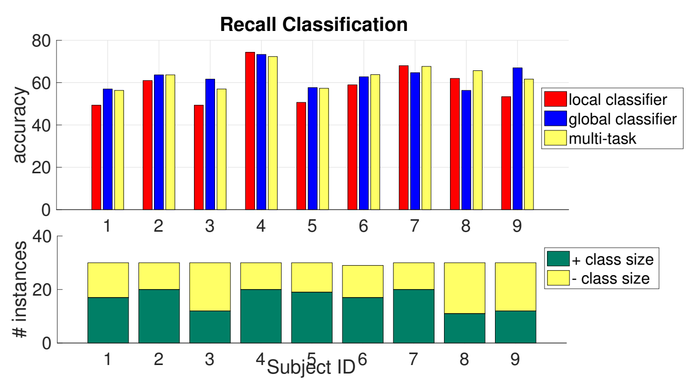

Figure 7: Recall level prediction results for individual subjects with 80% training set size.

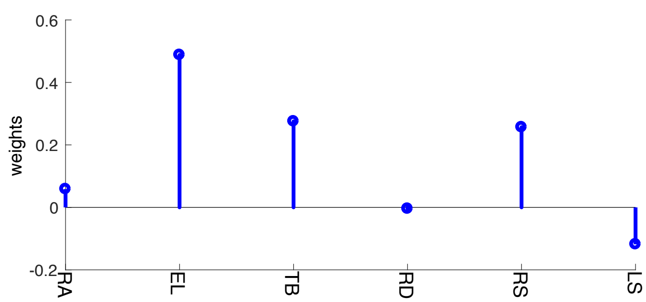

Figure 8: Feature weights for recall level prediction using global model. RA: reading concentration; EL: English level; TB: technical background; RD: reading duration; RS: reading speed; LW: Linsear Write index

5 CONCLUSIONS

This paper investigates the feasibility of using EEG recordings to monitor students' cognitive states during their learning process. Features are extracted from a broad spectrum of EEG band waves to obtain a more comprehensive view of the brain signals. Our results showed that the concentration level and reading speed of the subjects can be accurately determined from their EEG signals alone. However, inferring the recall level is considerably harder, requiring additional side information particularly about the background knowledge of subjects. We also demonstrated the advantages of using a broad spectrum of frequency bands to derive features for the classification models. Finally, we showed that the choice of classifier has a significant impact on the model's performance.

ACKNOWLEDGMENTS

This research was supported in part by the National Science Foundation under Grant No. IIS-1615612.

REFERENCES

[1] 2015. Muse Developers. (2015). http://developer.choosemuse.com/research-tools
[2] 2017. Wikipedia:Linsear Write. (2017). https://en.wikipedia.org/wiki/Linsear_Write
[3] Chris Berka, Daniel J Levendowski, Michelle N Lumicao, Alan Yau, Gene Davis, Vladimir T Zivkovic, Richard E Olmstead, Patrice D Tremoulet, and Patrick L Craven. 2007. EEG correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation, space, and environmental medicine* 78, Supplement 1 (2007), B231–B244.
[4] Gary Garcia-Molina, Tsvetomira Tsoneva, and Anton Nijholt. 2013. Emotional brain–computer interfaces. *International journal of autonomous and adaptive communications systems* 6, 1 (2013), 9–25.
[5] Brahim Hamadicharef, Haihong Zhang, Cuntai Guan, Chuanchu Wang, Kok Soon Phua, Keng Peng Tee, and Kai Keng Ang. 2009. Learning EEG-based spectral-spatial patterns for attention level measurement. In *Circuits and Systems, 2009. ISCAS 2009. IEEE International Symposium on.* IEEE, 1465–1468.
[6] Marc W Howard, Daniel S Rizzuto, Jeremy B Caplan, Joseph R Madsen, John Lisman, Richard Aschenbrenner-Scheibe, Andreas Schulze-Bonhage, and Michael J Kahana. 2003. Gamma oscillations correlate with working memory load in humans. *Cerebral cortex* 13, 12 (2003), 1369–1374.
[7] Wolfgang Klimesch, HANNES Schimke, and Gert Pfurtscheller. 1993. Alpha frequency, cognitive load and memory performance. *Brain topography* 5, 3 (1993), 241–251.
[8] Nicole Landi. 2010. An examination of the relationship between reading comprehension, higher-level and lower-level reading sub-skills in adults. *Read Writ.* 23, 6 (2010), 701–717.
[9] X. Li, Q. Zhao, B. Hu, L. Liu, H. Peng, Y. Qi, C. Mao, Z. Fang, and Q. Liu. 2010. Improve Affective Learning with EEG Approach. *Computing and Informatics* 29, 4 (2010), 557–570. http://www.sav.sk/index.php?lang=en&charset=ascii&doc=journal&part=list_articles&journal_issue_no=11112385#abstract_5955
[10] Yongchang Li, Xiaowei Li, Martyn Ratcliffe, Li Liu, Yanbing Qi, and Quanying Liu. 2011. A real-time EEG-based BCI system for attention recognition in ubiquitous environment. In *Proceedings of 2011 international workshop on Ubiquitous affective awareness and intelligent interaction.* ACM, 33–40.
[11] Ning-Han Liu, Cheng-Yu Chiang, and Hsuan-Chin Chu. 2013. Recognizing the degree of human attention using EEG signals from mobile sensors. *Sensors* 13, 8 (2013), 10273–10286.
[12] Axel Mecklinger, Arthur F Kramer, and David L Strayer. 1992. Event related potentials and EEG components in a semantic memory search task. *Psychophysiology* 29, 1 (1992), 104–119.
[13] Jon D Morris. 1995. Observations: SAM: the Self-Assessment Manikin; an efficient cross-cultural measurement of emotional response. *Journal of advertising research* 35, 6 (1995), 63–68.
[14] Dan Nie, Xiao-Wei Wang, Li-Chen Shi, and Bao-Liang Lu. 2011. EEG-based emotion recognition during watching movies. In *Neural Engineering (NER), 2011 5th International IEEE/EMBS Conference on.* IEEE, 667–670.
[15] Mark A Schier. 2000. Changes in EEG alpha power during simulated driving: a demonstration. *International Journal of Psychophysiology* 37, 2 (2000), 155–162.
[16] Daniel Szafir and Bilge Mutlu. 2012. Pay attention!: designing adaptive agents that monitor and improve user engagement. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* ACM, 11–20.
[17] D Wood. 2006. Modeling the relationship between oral reading fluency and performance on a statewide reading test. *Educational Assessment* 11 (2006), 85–104.
[18] Jiayu Zhou, Jianhui Chen, and Jieping Ye. 2012. MALSAR: Multi-task learning via structural regularization [Online]. http://www.MALSAR.org. (2012).