Confusion detection using cognitive ability tests

Caroline Dakoure[1], **Mohamed Sahbi Benlamine**[1], **Claude Frasson**[1]
[1] Département d'Informatique et de Recherche Opérationnelle, Université de Montréal
2920 Chemin de la tour, Montréal (QC), Canada, H3T 1J4
{caroline.dakoure, ms.benlamine}@umontreal.ca, frasson@iro.umontreal.ca

Abstract

It is of great importance to detect users' confusion in a variety of situations such as orientation, reasoning, learning, and memorization. Confusion affects our ability to make decisions and can lower our cognitive ability. This study examines whether a confusion recognition model based on EEG features, recorded on cognitive ability tests, can be used to detect three levels (low, medium, high) of confusion. This study also addresses the extraction of additional features relevant to classification. We compare the performance of the K-nearest neighbors (KNN), support vector memory (SVM), and long short-term memory (LSTM) models. Results suggest that confusion can be efficiently recognized with EEG signals (78.6% accuracy in detecting a confused/unconfused state and 68.0% accuracy in predicting the level of confusion). Implications for educational situations are discussed.

Introduction

Confusion is a state where an individual does not understand what is going on, what they should be doing, what something means, who is someone, or something. People may experience confusion in various contexts, notably in learning (D'Mello et al. 2014). During the learning process, confusion often occurs when students do not understand new knowledge. Learners' emotions influence their learning experience. For example, when they are confused, it can lead to erroneous decision-making and affect their performance, engagement, and cognitive load. In learning, confusion may help increase engagement and deepen knowledge, or it may lead to frustration and boredom if there is no understanding after a certain amount of time. The intensity and duration of the confusion appear to be factors of frustration or boredom (Arguel et al. 2017). Confusion is a condition that is also very present in people with dementia (Berry 2014) because of the decline in their cognitive abilities. Early detection of this state may help better understand a person's behavior in a given situation and thus treat or help them more effectively. Confusion can have an immediate negative impact on all of us, as in the case of confused drivers who crash (Beanland et al. 2013). Therefore, detecting confusion offers many benefits.

Emotion detection techniques based on brain activity are becoming increasingly popular. Electroencephalography (EEG), for example, can detect emotions using electrodes placed on the scalp. The electrodes record the brain's electrical activity that comes from the excitation of neurons that receive and transmit information. We hypothesize that we can create an effective model to detect three levels of confusion (low, medium, high confusion) with EEG signals.

There are a variety of cognitive exercises that generate confusion (Zhou et al. 2018). Therefore, we decided to conduct an experiment with cognitive ability tests to generate confusion and record it. We recorded the EEG signals of ten participants when they solved five series of different cognitive exercises. We preprocessed the EEG signals using MATLAB and EEGLAB. We extracted the power spectra from these cleaned EEG signals and used them as input to train our models. We used the levels of confusion self-reported by participants as outputs from our learning models.

The paper is organized as follows. We introduce the literature review of confusion recognition. Then we describe the research methodology of our experiment. In the next section, we present the dataset. Then we introduce the models and metrics we used, and we show the results. Finally, we discuss the results and conclude our study.

Confusion recognition

Literature review

Several experiments used EEG signals to detect two-level confusion (confused/unconfused). In recent years, studies have mainly focused on recognizing confusion to improve learning. In 2018, the brain activity of sixteen participants solving Raven's progressive matrices (Raven 2000) was recorded (Zhou et al. 2018). The best model was a

Copyright © 2021, by the authors. All rights reserved.

Convolutional Neural Network that achieved 71.36% accuracy. In the same year, a Bidirectional Long Short-Term Memory reached 75% accuracy in detecting confusion in ten adults watching massive open online course videos (Wang, Wu, and Xing 2018). Confusion is also associated with electroencephalography in the medical field. Mental confusion in 174 patients was detected using Convolutional and Recurrent Neural networks (Sun et al. 2019). The findings showed that it was possible to continuously track mental confusion in the intensive care unit, despite the model's accuracy not being reported. In 2020, one study investigated whether it was possible to find COVID-19-specific patterns from EEG signals of 23 confused patients (Petrescu, Taussig, and Bouilleret 2020). This study found EEG alterations in less than half of the participants. The detection of confusion is also possible by combining EEG signals with other sources. By mixing EEG signals with video features, the Sedmid model achieved 87.8% accuracy (Yang et al. 2016).

Our contribution

In this study, we performed multinomial classification to detect several levels of confusion. We did not only predict whether an individual is confused or not but also how confused they are.

Research methodology

Equipment

Emotiv Epoc

The Emotiv Epoc is a portable neuroheadset device with 16 electrodes (14 channels and 2 references behind the ears). The electrodes are positioned at AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4. The EEG signals are in μV with a sampling rate of 128 samples/s and frequencies' range between 0.2-43 Hz.

Emotiv Xavier Testbench

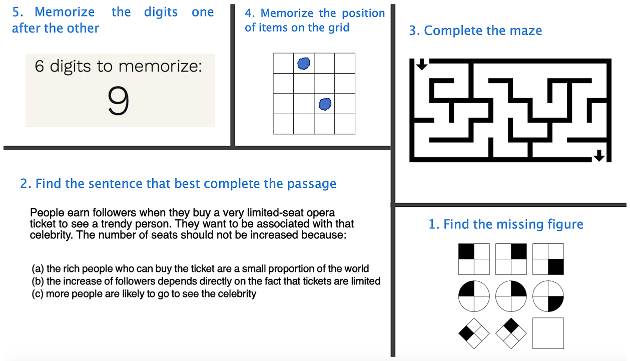
Emotiv Testbench is a software that receives the EEG signals collected by Emotiv Epoc. It displays the EEG signals in μV in real-time and has an option to save them. It also shows the quality of the sensors' contact on the scalp: green for good connection, orange for medium connection, and red for poor connection.

Experimental protocol

We recruited ten undergraduate students (5 women, 5 men) from the Department of Computer Science and Operations Research of the University of Montreal to participate in our experiment. They ranged in age from 22 to 33 (mean = 27.1, STD = 3.7) years old. As the Science and Health Research Ethics Board of the University of Montreal approved this research, all participants signed a consent form before beginning the experiment. Then the experimenter had the participants sit on a chair and checked the chair to maintain a good view on the computer screen. Next, the experimenter installed the Emotiv Epoc following the international 10-20 system to ensure reproducibility. Before the participants started the exercises, the experimenter ran Emotiv TestBench and checked the sensors' contact on the scalp. The experimenter then started to display and record the EEG data on the Emotiv TestBench, and the participants started to solve the cognitive ability tests. The cognitive ability tests were developed at the Heron lab of the University of Montreal. The experimenter also asked each subject to minimize body movements during recording to reduce noise on the collected data. Once they finished the exercises, participants were compensated $20 for participating and debriefed at the end. Each participant participated only once, and the experiment session lasted approximately one hour.

Cognitive ability tests

Fig. 1 introduces five series of exercises that we developed based on the following recognized cognitive tests: Raven's progressive matrices (Raven 2000), Gmat critical reasoning test (Kuncel, Credé, and Thomas 2007), WAIS-IV (Benson, Hulac, and Kranzler 2010). In the first series, the participants had to select the missing figure from a set of geometric figures. In the second series of exercises, the participants read a short text and choose the statement that best completes the passage. In the third series, the participants completed 2d mazes. In the fourth series, the participants memorized the position of items on a grid. Finally, in the last series, the participants had to memorize a sequence of numbers. Each set of exercises contained an example and four exercises with different levels of increasing difficulty resulting in 20 exercises. At the end of each exercise, the participants had to indicate their level of confusion (no confusion, slightly confused, moderately confused, very confused).


5. Memorize the digits one after the other
6 digits to memorize:
9
4. Memorize the position of items on the grid
3. Complete the maze
2. Find the sentence that best complete the passage
1. Find the missing figure

People earn followers when they buy a very limited-seat opera ticket to see a trendy person. They want to be associated with that celebrity. The number of seats should not be increased because:

(a) the rich people who can buy the ticket are a small proportion of the world
(b) the increase of followers depends directly on the fact that tickets are limited
(c) more people are likely to go to see the celebrity

Figure 1: Developed cognitive ability tests

Dataset

Our initial EEG signals

For each participant, using the Emotiv Testbench, we obtained an EDF file with the EEG data for all the 20 exercises done. At each second, 128 EEG signals were recorded for all 14 signals.

Preprocessing with EEGLAB

To make sense of the EEG signals and subsequently extract meaningful measurements, we preprocessed the data with MATLAB r2020b and EEGLAB v2020.0. We first observed our entire dataset and deleted the exercises where most of the data was noise. We then used a high-pass filter at 0.5Hz. The reason for using the high-pass filter is that we wanted to use ICA (Comon 1994), and ICA is sensitive to low frequencies. The Emotiv Epoc is not supposed to record above 43Hz, so we applied a low-pass filter at 43Hz. After filtering the data, we examined the data and removed the artifacts. The artifacts often have a higher amplitude than the brain signals. Typical examples of artifacts are high-frequency artifacts such as muscles and low-frequency artifacts such as eye movements. Other examples include electrical noise, a discontinuity in the signal, and so on. After removing the artifacts, we ran the ICA algorithm. ICA breaks down the signal into independent components. We had 14 electrodes, each recording a signal from different brain sources (blink, jaw movement, physiological rhythm, etc.). The electrodes contained various sources. With ICA we obtained the signal from a particular source, for example, the blink of an eye. We could, therefore, reject unnecessary components (muscles, eye movements, etc.) that we had not previously removed.

From EEG signals to the power spectrum

We decided to look at the different frequency bands, which are often used to portray brain activity, to avoid long training. These bands represent the speed at which the brain processes information and interacts with other areas of the brain. Frequency can be defined as the number of times a phenomenon occurs in a given time. A high frequency will, therefore, indicate a phenomenon that occurs frequently and vice versa. The principle of switching from the time domain to the frequency domain is to decompose the signal into several periodic signals to get a different view of the signal and see the frequencies. Having access to the frequencies makes it possible to see how many times each amplitude occurred. An amplitude that occurs frequently will have a higher frequency than an amplitude that occurs more rarely. Fig. 2 illustrates the transition from time to frequency domain.

We first decomposed the EEG signals into the frequency domain using the Fast Fourier Transform (FFT). After getting the FFT, we computed the average band power (Welch 1967). The average band power is a number in $\mu V^2 Hz^{-1}$ summarizing the contribution of the frequency band to the total signal power. We computed the average band power for the following bands (Marzbani, Marateb, and Mansourian 2016): delta (1-4 Hz), theta (4-8 Hz), alpha (8-13 Hz), beta (13-30 Hz), and gamma (30-43 Hz). Finally, the size of the input was 5007x14x5.

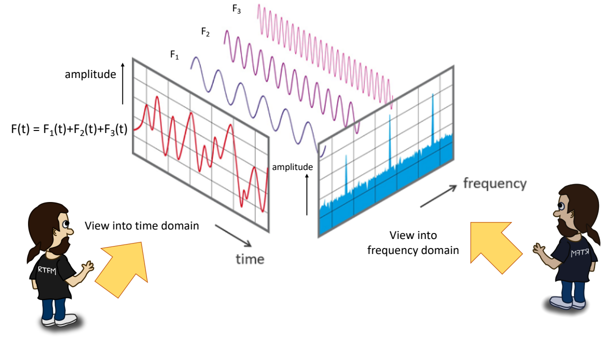


Figure 2: From time domain to frequency domain

Our output labels

After each cognitive exercise, participants self-reported their level of confusion. They could select the following: not confused, slightly confused, moderately confused, and very confused. We converted these confusion levels into a three-point Likert scale (low, medium, high confusion) and two options scale (no confusion, confusion). Initially, we had confusion levels associated with exercises of different durations, but we wanted to obtain a model that gave us the confusion at each second. We split the exercises into seconds and assigned the self-reported confusion level for the entire exercise to each second.

Dataset visualization

Fig. 3 shows a 2D visualization of our dataset with three levels of confusion. We obtained this visualization using sklearn.decomposition.PCA.

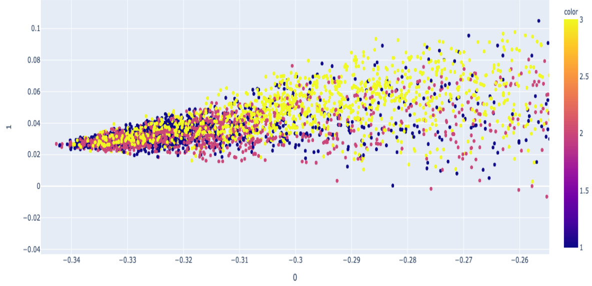


Figure 3: 2d visualization of the dataset. Low confusion is in blue, medium confusion in purple and high confusion in yellow.

Models and metrics

We trained 3 models for multinomial classification: K-nearest neighbors (KNN), a support vector memory (SVM), and a long short-term memory (LSTM). We split the data randomly using 5-fold cross-validation. We used SMOTE (Chawla et al. 2002) oversampling of the minority classes to balance classes in the training data. To evaluate our models, we wanted to take into account correct predictions and penalize incorrect predictions. We also had several classes, and we wanted them all to have the same importance. To meet the above requirements, we computed the subset accuracy. The subset accuracy measures the percentage of inputs in a subset that exactly matches the labels. For example, among the signals that the model associates with an unconfused state, it is the percentage of truly unconfused signals. Here is the formula for subset accuracy:

$$SubAcc = \frac{1}{n} I(y_i = z_i) \tag{1}$$

where n is the number of samples of the class, e.g., the total number of unconfused signals. y is the true level of confusion associated with the signal of index i. z is the predicted level of confusion for the signal of index i. I is the indicator function which returns 1 if y = z and 0 otherwise. It is also a widely used metric as it is one of the standard metrics of the Scikit-learn library for multiclass classification. In the following paragraphs, we used the word "accuracy" instead of "subset accuracy" for visibility reasons.

KNN

The KNN algorithm is a non-parametric algorithm that we used for the classification. We chose it because it does not need a training period and it only has a few parameters to tune. It is a fast and easy to implement algorithm resulting in quick first results. KNN classifies a given data point according to the majority of its k closest neighbors.

We implemented KNN with the sklearn.neighbors.KNeighborsClassifier class. To tune the k hyperparameter, we used the class sklearn.model_selection.GridSearchCV. The Grid Search tests a given set of values for each specified hyperparameter and gives the model's accuracy at each test. We tested values of k between 1 and 20.

SVM

After viewing our dataset with PCA, we saw that it was not linearly separable. We used a support vector machine (SVM) because it is a robust algorithm that can be applied with non-linear data and has often been used in the literature to classify emotions (Atkinson and Campos 2016). Moreover, it is efficient in high-dimensional space. The SVM algorithm uses a technique called kernel trick to transform the data. It then separates the data according to their classes.

We implemented SVM with the sklearn.svm.SVC class. To tune the regularization parameter C and the kernel coefficient $gamma$ we used the Grid Search. We tested values for C between [0.01, 100] and values for $gamma$ between [0.01, 100].

LSTM

The long short-term memory (LSTM) neural network can learn over long sequences to predict the next one. It is often used with time series (Lipton et al. 2015) because of its memory capacity and therefore looked promising with our EEG signals. We chose it as a third model to compare it with the others' results.

We implemented LSTM with the tf.keras.layers.LSTM class. To tune the number of neurons and the number of epochs, we used the Grid Search. We tested values for the number of neurons between [35,48] and values for the number of epochs between [500,1000].

Results

In this last section, we compare the results of the classification of our KNN, SVM, and LSTM. We also compare our results with those of state-of-the-art. We obtained all our scores with a random 5-fold cross-validation and subset accuracy metrics.

We initially focused on the multinomial classification of levels of confusion. We started with the KNN model, which has a good time complexity allowing us to have quick first results. We tested KNN with the Euclidean distance and a neighbor number between 1 and 20. KNN obtained the best accuracy of 65.3% with $k = 20$ as shown in Fig. 4.

We then switched to the SVM model as it has been used in the literature for EEG detection (Atkinson and Campos 2016). We evaluated SVM with different hyperparameters as described in Table 1. The kernel coefficient $gamma$ seems to be the most significant parameter for accuracy with the SVM model. The regularization parameter C has also allowed to improve the accuracy in a consequent way. We got the best accuracy of 68.0%, using $gamma$ = 100 and C = 100. Fig. 5 shows the confusion matrix of the SVM.

We then trained an LSTM because it is known for its performance with the times series (Lipton et al. 2015). Another advantage is that it learns patterns from data on the contrary of KNN that compute distances. We configured the number of neurons, and the number of epochs. We got the best accuracy of 65.1% using 4 hidden layers, 47 hidden units, a batch of size 128, and 1000 epochs as shown in Table 2.

After making a multinomial classification of the levels of confusion, we wanted to see our models' accuracy for a simplified task: the detection of the confused or unconfused state. Following the same pipeline as for the multinomial classification, the SVM obtained the best accuracy of 78.6%. Fig. 6 shows its confusion matrix.

Finally, we compared our models with state-of-the-art in Table 3. Our best model for detecting the state of confusion achieved an accuracy of 78.6%, placing it second to state-of-the-art.

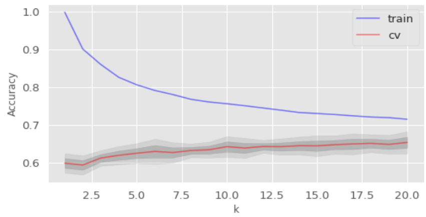


Figure 4: Validation curve on the KNN parameter k for detecting 3 levels of confusion. For k=20, the accuracy is 65.3%.

C	gamma	5-fold accuracy
0.01	100	55.0 ± 1.4
1	100	64.8 ± 1.0
10	100	67.7 ± 0.8
100	100	68.0 ± 1.1
10	1	57.9 ± 1.4
10	10	64.2 ± 1.3

Table 1: Accuracy of SVM model for detecting three levels of confusion.

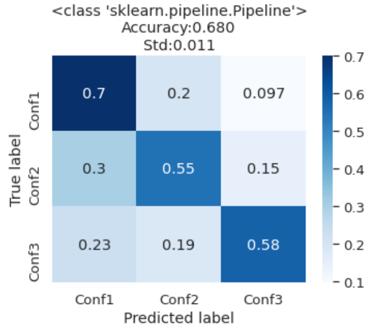


Figure 5: Confusion matrix of SVM model for detecting three levels of confusion.

Number of epochs	Number of neurons	Accuracy
500	35	62.0
1000	46	63.6
1000	47	65.1
1000	48	64.7

Table 2: Accuracy of the LSTM to detect three levels of confusion with a batch_size of size 128 and 4 hidden layers.

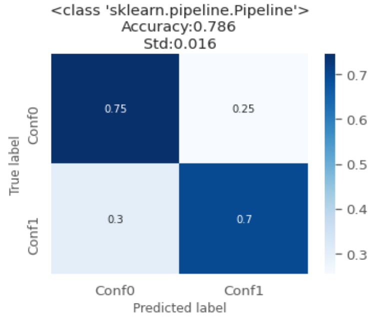


Figure 6: Confusion matrix of SVM model for detecting confused/unconfused state of confusion

Method	Features	Detection	Accuracy
Yang et. al. (2016)	EEG signals, audio-visual sources	Confused, unconfused	87.8
Our SVM	Power spectrum	Confused, unconfused	78.6
Wang et al. (2018)	Power spectrum, attention, meditation, EEG signals	Confused, unconfused	75.0
Our SVM	Power spectrum	Low, medium, high confusion	68.0

Table 3: Overview of best classifiers accuracy

Discussion

We hypothesized that we could detect three levels of confusion, which has not been seen in state-of-the-art. Our best model achieved 68.0% accuracy. Given the accuracy achieved by the two-level (confused/unconfused) confusion recognition models in state-of-the-art, confusion is a complex emotion that seems difficult to recognize. More than two-level emotion recognition models generally have less accuracy than two-level models (Jun, and Smitha 2016). Hence, for our three-level model, an accuracy of 68% seems to be a promising result. We also wanted to develop a model to detect two-level confusion and compare this model with state-of-the-art. Our best model for detecting the state of confusion used fewer features than state-of-the-art models. Despite this, it ranked second with an accuracy of 78.6% (Table 3). Its advantage is that it is easier to use in real-time as it only uses the features given by the neuroheadset. The KNN, SVM, and LSTM algorithms achieved close accuracy. The question then arises as to which one to choose. KNN computes the distances and, therefore, does not find patterns in the data. The only information KNN gives is that the training data form clusters and that the examples of the same class are close in the feature space. KNN, therefore, requires representative training samples because it cannot abstract and learn patterns. If the data is exposed to certain transformations that change distances, KNN can lose its efficiency. Another of its constraints is that it always needs all the data to make a new prediction. On the other hand, SVM and neural networks learn patterns on the data. Thus, they can be more appropriate for real-time data that may be isolated from the other clusters. SVM has the advantage of not requiring a large number of training examples.

Our confusion recognition model can be used in real-time to assess an individual's confusion in different applications such as education or health. The requirement is to have an EEG device that records brain activity. EEG devices have advantages, such as the fact that they can be used with people who cannot make a motor response, but they also have limitations. Their most known drawback is their low spatial resolution that creates noise. We reduced this noise during preprocessing, but it is still present. EEG also requires careful placement of electrodes, and hair, skull shape, user movement, can make detection difficult. All these factors limit the use of these technologies in clinical settings. Therefore, our model can be implemented in real-world scenarios, keeping in mind that it provides help but does not have ultimate reliability and should be used in applications where detection reliability is not critical.

Conclusion

The recognition of confusion is important in adapting learning, care, or broadly a system, to a user. This study demonstrates the possibility of multiclass classification of confusion for three levels of intensity. In addition, our best model for classifying three levels of confusion reached 68.0% accuracy. We also predicted the confused/unconfused state with an accuracy of 78.6%. It would be interesting for a future study to analyze whether confusion led to engagement or frustration and find a way to predict if confusion is likely to have a positive or negative outcome. The neurological activities related to emotions can be very different from one person to another. Therefore, it would also be desirable for our model to be adaptable to the participants.

Acknowledgments

We acknowledge NSERC-CRD, BMU and Prompt for funding this research.

References

Arguel, A.; Lockyer, L.; Lipp, O. V.; Lodge, J. M.; and Kennedy, G. 2017. Inside Out: Detecting Learners' Confusion to Improve Interactive Digital Learning Environments. *Journal of Educational Computing Research* 55 (4): 526–551.

Atkinson, J.; and Campos, D. 2016. Improving BCI-Based Emotion Recognition by Combining EEG Feature Selection and Kernel Classifiers. *Expert Systems with Applications* 47: 35–41.

Beanland, V.; Fitzharris, M.; Young, K. L.; and Lenné, M. G. 2013. Driver Inattention and Driver Distraction in Serious Casualty Crashes: Data from the Australian National Crash In-Depth Study. *Accident Analysis & Prevention* 54 (May): 99–107.

Benson, N.; Hulac, D. M.; Kranzler, J. H. 2010. Independent Examination of the Wechsler Adult Intelligence Scale—Fourth Edition (WAIS-IV): What Does the WAIS-IV Measure? *Psychological Assessment* 22 (1): 121.

Berry, B. 2014. Minimizing Confusion and Disorientation: Cognitive Support Work in Informal Dementia Caregiving. *Journal of Aging Studies* 30: 121–30.

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. SMOTE: Synthetic Minority over-Sampling Technique. *Journal of Artificial Intelligence Research* 16: 321–57.

Comon, P. 1994. Independent Component Analysis, a New Concept? *Signal Processing* 36 (3): 287–314.

D'Mello, S.; Lehman, B.; Pekrun, R.; and Graesser, A. 2014. Confusion Can Be Beneficial for Learning. *Learning and Instruction* 29: 153–70.

Jun, G.; and Smitha, K. G. 2016. EEG based stress level identification. In *2016 IEEE international conference on systems, man, and cybernetics (SMC)*: 003270–003274.

Kuncel, N. R.; Credé, M.; and Thomas, L. L. 2007. A Meta-Analysis of the Predictive Validity of the Graduate Management Admission Test (GMAT) and Undergraduate Grade Point Average (UGPA) for Graduate Student Academic Performance. *Academy of Management Learning & Education* 6 (1): 51–68.

Lipton, Z. C.; Kale, D. C.; Elkan, C.; and Wetzel, R. 2015. Learning to Diagnose with LSTM Recurrent Neural Networks, *November. arxiv.org/abs/1511.03677v7* .

Marzbani, H.; Marateb, H. R.; and Mansourian, M. 2016. Neurofeedback: A Comprehensive Review on System Design, Methodology and Clinical Applications. *Basic and Clinical Neuroscience* 7 (2): 143.

Petrescu, A.; Taussig, D.; and Bouilleret, V. 2020. Electroencephalogram (EEG) in COVID-19: a systematic retrospective study. *Neurophysiologie Clinique* 50(3): 155–165.

Raven, J. 2000. The Raven's Progressive Matrices: Change and Stability over Culture and Time. *Cognitive Psychology* 41 (1): 1–48.

Sun, H.; Kimchi, E.; Akeju, O.; Nagaraj, S. B.; McClain, L. M.; Zhou, D. W.; Boyle, E.; Zheng, W.; Ge, W.; and Westover, M. B. 2019. Automated tracking of level of consciousness and delirium in critical illness using deep learning. *NPJ digital medicine* 2(1): 1-8.

Wang, H., Wu, Z.; and Xing, E. P. 2018. Removing Confounding Factors Associated Weights in Deep Neural Networks Improves the Prediction Accuracy for Healthcare Applications. In *Biocomputing 2019*, 54–65.

Welch, P. 1967. The Use of Fast Fourier Transform for the Estimation of Power Spectra: A Method Based on Time Averaging over Short, Modified Periodograms. *IEEE Transactions on Audio and Electroacoustics* 15 (2): 70–73.

Yang, J., Wang, H.; Zhu, J.; and Xing, E. P. 2016. SeDMiD for Confusion Detection: Uncovering Mind State from Time Series Brain Wave Data, *November. arxiv.org/abs/1611.10252v1* .

Zhou, Y.; Xu, T.; Li, S.; and Li, S. 2018. Confusion State Induction and EEG-Based Detection in Learning. In *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*: 3290–93.