Journal of Neural Engineering

PAPER

Confused or not: decoding brain activity and recognizing confusion in reasoning learning using EEG

Journal of Neural Engineering

RECEIVED
11 November 2022

REVISED
18 February 2023

ACCEPTED FOR PUBLICATION
28 February 2023

PUBLISHED
17 March 2023

PAPER

Confused or not: decoding brain activity and recognizing confusion in reasoning learning using EEG

Tao Xu[1], Jiabao Wang[1], Gaotian Zhang[1], Ling Zhang[2] and Yun Zhou[2,*]

[1] Northwestern Polytechnical University, School of Software, Xi'an, People's Republic of China
[2] Faculty of Education, Shaanxi Normal University, Xi'an, People's Republic of China
[*] Author to whom any correspondence should be addressed.

E-mail: zhouyun@snnu.edu.cn

Keywords: emotion recognition, EEG, confusion, affective computing

Abstract

Objective. Confusion is the primary epistemic emotion in the learning process, influencing students' engagement and whether they become frustrated or bored. However, research on confusion in learning is still in its early stages, and there is a need to better understand how to recognize it and what electroencephalography (EEG) signals indicate its occurrence. The present work investigates confusion during reasoning learning using EEG, and aims to fill this gap with a multidisciplinary approach combining educational psychology, neuroscience and computer science. *Approach.* First, we design an experiment to actively and accurately induce confusion in reasoning. Second, we propose a subjective and objective joint labeling technique to address the label noise issue. Third, to confirm that the confused state can be distinguished from the non-confused state, we compare and analyze the mean band power of confused and unconfused states across five typical bands. Finally, we present an EEG database for confusion analysis, together with benchmark results from conventional (Naive Bayes, Support Vector Machine, Random Forest, and Artificial Neural Network) and end-to-end (Long Short Term Memory, Residual Network, and EEGNet) machine learning methods. *Main results.* Findings revealed: 1. Significant differences in the power of delta, theta, alpha, beta and lower gamma between confused and non-confused conditions; 2. A higher attentional and cognitive load when participants were confused; and 3. The Random Forest algorithm with time-domain features achieved a high accuracy/F1 score (88.06%/0.88 for the subject-dependent approach and 84.43%/0.84 for the subject-independent approach) in the binary classification of the confused and non-confused states. *Significance.* The study advances our understanding of confusion and provides practical insights for recognizing and analyzing it in the learning process. It extends existing theories on the differences between confused and non-confused states during learning and contributes to the cognitive-affective model. The research enables researchers, educators, and practitioners to monitor confusion, develop adaptive systems, and test recognition approaches.

1. Introduction and Background

The state of cognitive disequilibrium is characterized by an inconsistency between an individual's cognitive schema and the new information [1–3]. It is closely tied to learning activities that involve absorbing knowledge, reasoning, or solving problems [4–6]. As a result of cognitive disequilibrium, students are fostered to think deeply and gain a more profound understanding of the learning subjects and materials. Confusion is an important emotional indicator of cognitive disequilibrium [1, 7]. Students feel confused when their cognitive structures are not consistent with the new information, or when they seek rules in logical reasoning or solving problems, but cannot progress further [8–10]. According to the model of affect dynamics [1], confusion is a primary epistemic emotion that correlates to learners' engagement, frustration, and boredom. Furthermore, confusion is more common in learning than other emotions [5, 7, 11]. Although confusion is an unpleasant emotion, the behavior of resolving confusion during a controllable period has been proved to be beneficial for learning [12–14], fostering students to engage highly in learning activities. To improve students' deep engagement and learning outcomes, researchers integrate specific tasks in games and educational applications to manage optimal confusion [5, 15]. However, the study of confusion in learning is still in its infancy. More research is required to determine what features can represent confusion, and what brain functions are associated with this epistemic emotion.

Recognizing, monitoring, and analyzing confusion is the prerequisite of academic emotion regulation [16] and adaptive learning intervention, offering an opportunity to improve learning experience and outcomes. Researchers have measured confusion using various methods, including self-reports [12], facial expressions [17], dialogues [18], eye-tracking [19], and physiological measures [6, 20]. Electroencephalography (EEG) is a physiological and objective measure that has three-fold advantages over other methods for recognizing emotions [21–23]. First, EEG directly reflects the intrinsic mental states of human beings. Second, it discloses the variation of mental states over time due to its good temporal resolution in nature. Third, only a few simple preparations are required to obtain accurate brainwave data through portable EEG [24] compared with other physiological methods. Research on confusion in learning is still in its early stages. On the one hand, most current studies do not go beyond engagement and workload [25, 26]. On the other hand, a literature search revealed that there are few studies to date on the monitoring and analysis of confusion using EEG signals. For example, Liang *et al* studied alpha band changes of confusion states induced by problem-solving and reasoning [27]. Reñosa *et al* used the power spectrum of all EEG frequencies as features and artificial neural networks (ANNs) to classify low, medium and high levels of confusion [28]. Zhou *et al* [6] used a commercially available device to collect EEG data and proposed an end-to-end method to classify two states: confused and non-confused, using a labeling technique based on participants' self-assessment. However, there is still a lack of research on what the EEG shows during confusion. Additionally, compared to other emotions, no previous research has built annotated confusion datasets or databases to serve as a benchmark for the development of recognition methods. Therefore, there is an urgent need to build databases for recognizing and analyzing students' confusion during learning.

To decode brain activity associated with confusion and recognize this emotion using EEG and achieve breakthroughs, many disciplinary approaches are required [29]. For instance, establishing elicitation techniques and adopting an appropriate theoretical framework of emotion require psychology. Understanding how affective emotions are represented in the EEG is made easier by neuroscience. The most promising field for obtaining reliable features for affective computing as well as accurately and constantly interpreting affective state is information technology. As a result, the present work fills the gap on confusion in learning by combining approaches from educational psychology, neuroscience, and computer science.

In this work, we design and conduct an experiment to invoke confusion using logical reasoning tests based on the [6]. Inferential learning is one of the most essential parts of learning, referring to the learning which enables people to construct new knowledge by thinking [30]. When people are reasoning and struggle to find a solution, they experience cognitive disequilibrium and feel confused. We then propose a subjective and objective joint labeling technique to categorize states related to confusion, and provide an EEG database with benchmark results for recognizing and analyzing this epistemic emotion. The results showed that the power in the delta, theta, alpha, beta, and lower gamma bands significantly differed between the confused and non-confused conditions. This indicates that learning targets requires more attention and cognitive resources when an individual is confused. We also observed a higher cognitive load and attention when people felt confused by analyzing theta and alpha power. We provide CAL database—an EEG database for **C**onfusion **A**nalysis in **L**earning. CAL focuses on the emotion connected to cognitive activities occurring in learning. Furthermore, we extracted six features from the frequency, spatial, and temporal domains, and then conducted binary classification (confused, non-confused) and four-class classification (confused, non-confused, think-right, and guess) on subjects-dependent and subject-independent tasks respectively. We evaluated three end-to-end and four conventional machine learning methods on accuracy and F1 score. These comparisons are served as a benchmark for developing recognition methodologies and educational tools, and lays the groundwork for future efforts.

The layout of the paper is as follows. In section 2, we review the role of confusion in learning, cognitive states and band power variations, EEG-based emotion detection methods, and datasets/databases for emotion analysis. We present the stimuli selection and experiment design in section 3 and the experiment setup in section 4. We illustrate confusion emotion analysis in section 5 and the methodology of confusion recognition in section 6. In section 7, we discuss the findings with the limitations and future research objectives. We conclude in the final section.

2. Relevant literature

2.1. The role of confusion in learning

Literally, confusion in learning is the feeling of being confused when absorbing knowledge or working on a problem. Existing studies attempt to depict confusion as an emotion or an affective state due to its properties shared with the affect. In the work [8], confusion is considered to be an affective state, emerging as a product of an individual's appraisals of relevant events. Some work identified confusion as an emotion, more accurately, a learning emotion [22], an academic emotion [31], or an epistemic emotion [14]. Despite the lack of a unified understanding on a definition of 'confusion' in learning and a lack of studies on its neural substrates, it has distinct properties, such as a connection to cognition and a dual-sided role in learning.

Confusion is identified as a transitional emotion [1] according to the model of confusion dynamics [8]. Confusion benefits or harms the learning due to this transition property [1, 8, 13, 32]. When the confusion is resolved, the observed model follows the cycle of engagement-confusion-engagement. Otherwise, there will be a confusion-frustration and frustration-disengagement transition [8]. Most recent works have verified this dual-sided role of confusion in learning. The key point determining whether confusion produces positive or negative outcomes is resolution. As a result, identifying confusion is a requirement for learning intervention, which provides an opportunity to improve the learning experience and outcomes.

Confusion is highly relevant to learning activities like reasoning or problem solving, which occurs when the inconsistency triggers a cognition disequilibrium resulting in uncertainty about how to move further [8, 13, 33]. Confusion represents disequilibrium in cognition. To overcome the cognitive disequilibrium and move forward, the individual should reflect on materials and process them at deeper levels of comprehension. As a result, once confused, the individual engages in profound thinking and deep learning.

The research on confusion in learning, particularly in the context of reasoning and problem-solving, is still in its infancy.

2.2. Cognitive states and band power variations

EEG records an electrogram of scalp electrical activity that represents macroscopic brain activity. Most EEG applications for educational purposes are non-invasive. The primary EEG frequencies consist of six wave patterns: delta (2–4 Hz), theta (4–8 Hz), alpha (8–12 Hz), beta (12–30 Hz), lower gamma (30–70 Hz), and upper gamma (70–150 Hz) [34]. But there are no clear lines that separate the bands. Research has shown that delta, theta, alpha, beta, and lower gamma are significantly associated to cognitive and affective states in humans, representing the performance of cognitive processes [34–39]. Therefore, one way to understand how the brain works during cognitive and emotional states is to look at EEG data and mine EEG patterns.

The most recent EEG studies on cognitive state and emotion for educational purposes focus on attention or engagement [40–44], cognitive load [45, 46] and some basic emotions like happiness and fear [22, 26]. For instance, the EEG-based brain-computer interfaces (BCI) in the PAY ATTENTION study [41] collected EEG at the FP1 site to track changes in attention. The adaptive agent robot employed visual and auditory cues like rhythmic hand raising to help students redirect their attention when it fell below the predetermined threshold. The proposed BCI has been proved to enhanced the learning performance. Theta band power is thought to increase with cognitive resource demand and this rise is most noticeable in the frontal-central regions [47–52]. In addition, it has been found that alpha band power decreases as task difficulty increases [53, 54]. However, the findings on how alpha band power varies with increased task difficulty or the number of concurrent tasks are mixed. Results from several studies on the alpha band revealed an increase in alpha band power that was similar to the theta band's pattern [40, 55, 56].

As the importance of confusion studies has been recognized by researchers, they have begun to investigate the band changes associated with confusion. For example, Liang *et al* examined alpha band changes of confusion states induced by problem solving and reasoning [27]. They found that confusion can cause more brain activity in the cortical regions associated with the tasks that cause confusion. They also found that the frontal region is associated with the processing of novel or unfamiliar information, and the parietal-temporal regions are involved in sustained attention or reorientation during confusion induced by lack of information. However, there is still a lack of research on what the EEG shows during confusion and the relationship between attention and cognitive load as reflected in brain activity. Thus, in this paper, we investigate the band power differences between confused and non-confused conditions, and discuss their indications.

2.3. EEG-based emotion detection methods

Arguel *et al* review self-report, behavioral, and physiological methods for detecting confusion in digital learning environments [57]. Behavioral responses include facial expressions, postures, conversational cues, and learner–computer interaction. For example, Pachman *et al* used eye-tracking to detect confusion and explored the correlation of self-report and fixation data [19]. Physiological responses include electrodermal activity, heart rate and heart rate variability, brain imaging, and pupillometry. For example, Wang *et al* employed lecture clips to induce

Table 1. The comparison of the existing related datasets/databases and CAL.

Datasets/databases	Emotion categories	Emotional stimuli	Labeling method	Participant number
DEAP [63]	Arousal, valence and dominance	Music and video	Self-assessment	32
MAHNOB-HCI [64]	Neural, anxiety, amusement, sadness, joy, disgust, anger, surprise, and fear	Film clips	Self-assessment	27
SEED [65]	Positive, neutral and negative	Film clips	Self-assessment	15
HR-EEG4EMO [66]	Positive (amusement and tenderness) and negative (anger, fear, disgust, and sadness)	Film clips	Self-assessment	27
ASCERTAIN [67]	Affect (described in dimensional model-arousal and valence) and personality	Film clips	Self-assessment	58
AMIGOS [68]	Affect (basic emotions: neutral, disgust, happiness, surprise, anger, fear and sadness), personality, and mood	Film clips	Self-assessment	40
DREAMER [69]	Amusement, excitement, happiness, calmness, anger, disgust, fear, sadness and surprise	Film clips	Self-assessment	23
MPED [70]	Joy, funny, anger, fear, disgust, neutrality	Video clips	Self-assessment	30
CAL (This work)	Confused, non-confused, guess, think-right	Tests	Subjective and objective joint labeling	23

confusion and collect EEG for detection [58]. Among these methods, EEG has three advantages over others for recognizing emotions. It reflects human mental states directly and can reveal mental state changes over time. Furthermore, portable EEG requires few simple preparations to obtain accurate brainwave data.

Methods for detecting cognitive and affective states include the use of indices, as described in section 2.2, and machine learning techniques. Machine-learning techniques are novel and effective tools for EEG research and application. We synthesize readings on EEG data classification and divide current approaches into two types: conventional and end-to-end. In conventional classification, EEG signals are filtered in the time, frequency, and spatial domains to extract features. Then, these features are used to train a classifier. Due to its good performance on EEG data and suitability for small sample size, support vector machine (SVM) is one of the most popular algorithms for building classifiers in EEG-based brain-computer interfaces [22, 59]. An ANN is a machine learning model inspired by the structure and function of the human brain, which has been used for EEG classification. For example, based on the EEG data collected by Wang *et al* [20], Reñosa *et al* used the power spectrum of all the brain wave frequencies as features and ANNs to classify low, medium and high levels of confusion [28]. The end-to-end approach can build a classifier from raw EEG data without handcrafted features, which is useful when it is unclear what features to extract [60, 61]. It eliminates feature extraction and selection by using a single neural network. For example, Zhou *et al* [6] investigated the induction of confusion and the feasibility of detecting confusion using EEG. They used commercial device to collect EEG data and proposed an end-to-end method to classify two states: confused and non-confused. The data were labeled based on participants' self-assessment.

2.4. Datasets/databases for emotion analysis

Before detecting emotions, it is important to evoke them precisely. People react differently to the same materials, causing emotional differences. Most current research uses self-assessment to measure and reduce the discrepancy between the desired and evoked emotion. Pictures, sound or music clips, and video clips are the most used stimulus materials in the emotion analysis. In recent years, standard databases of movie clips to evoke basic emotions have been proposed and build [62], although the number is still limited. To trigger learning emotions, researchers have attempted to use tests, pedagogical contents, pictures, sounds, and courses video clips. For example, Lehman *et al* employed pedagogical contents to trigger emotions like confusion, frustration, anxiety, and curiosity in one-to-one expert tutoring sessions [11]. Wang *et al* used courses clips that were selected from the Coursera to trigger the confusion in learning [58]. In conclusion, a high-quality induction is essential to the success of high-quality EEG data for emotion recognition.

As shown in table 1, we listed the most recent related datasets and databases and compared them with CAL. In addition to using movie clips or videos as visual-audio stimuli, some studies have used the International Affective Picture System as visual stimulation and the International Affective Digitized Sound System as auditory stimulation [71–75]. Our focus is on public datasets as the CAL is publicly accessible. Therefore, we included only publicly available datasets or databases in table 1.

After exploring these databases and comparing their characteristics, we found that most available databases focus on the common emotions, described by dimensional or discrete emotion models. The stimuli in these databases are video clips, and methods of labeling data are mainly based on self-assessment, except one also uses external annotation by annotators. Labeling technique is the core to the quality of the data. Most existing studies, such as DEAP [63], MAHNOB-HCI [64], and SEED [65] adopted the self-assessment method to label, which is subjective. The cultural background and education level may affect the perception of stimuli materials and result in a nuanced emotion. Researchers have attempted to eliminate the deviation between the ground truth and the perception. However, it is inevitable to introduce human subjective error. An accurate labeling technique is a challenge for the experimental design.

To the best of our knowledge, there are no datasets or databases for investigating learning emotions, especially confusion from physiological signals, and no dataset or database using tests as stimuli. CAL is an EEG database that uses reasoning tests to evoke confusion.

3. Experiment design

This work focuses on the confusion that is evoked in inferential learning. The stimuli and experiment design should meet the following requirements:

(i) Ensure that the subject is involved in logic reasoning tasks, which are not related to knowledge and culture, and can be applied to everyone;
(ii) Ensure that the subject gets confused when performing some tasks.
(iii) Ensure that the EEG data are labeled accurately.
(iv) Ensure that the EEG data are collected with the least noise.

3.1. Stimuli selection

Although researchers attempt to evoke the emotion accurately, there still is a gap between the pre-assigned stimulus and the evoked emotions. To solve these two issues and obtain valid labeled data for CAL, we carefully designed the experiment and compared students' test results with their perceptions.

To meet the first requirement, we employ Raven's Progressive Matrices [76] as confusion stimuli to design the experiment. The Raven's test is a non-verbal psychological test used to measure human intelligence and abstract reasoning. The visual geometric design is independent of education and cultural background. It is suitable for subjects spanning an extensive range of ages and professions, making the subjects focus on inference. As shown in figure 1, it consists of 3 × 3 visual geometric matrix with a missing piece. The missing part is in the third row and third column. The task is to find the correct missing piece via logic inferring. Two assumptions hide in the matrix [77]: (1) shapes in a single row or column are related following some image transformation, and (2) parallel rows or columns share the same image transformation.

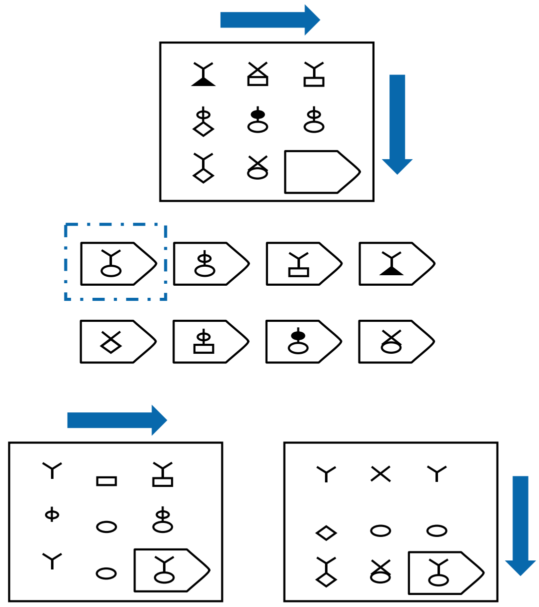


Figure 1. An example to illustrate how Raven's Progressive Matrices work.

3.2. Confusion induction

How to make sure the subject is confused for logical inference is the core issue in our experiment. Confusion occurs in the period from when people begin to work on the problem to before the time the problem is solved. We design an approach letting answering time be less than required. The difficulty of the test makes it impossible for most people to find a solution within a limited time. The method is to choose more complex test items and limit answering time.

Raven's Progressive Matrices consist of Raven's Standard Progress Matrices (SPM) and Raven's Advanced Progressive Matrices (APM). The problem-solving strategies of each group are similar, but the difficulty is increasing. SPM tests general abstract reasoning abilities such as visual discrimination, graphic imagination, comparison, reasoning, and series relationship. SPM is in the form of five groups of test items from easy to difficult: A, B, C, D, and E, and each group contains 12 test items. APM is designed for testing extraordinary intelligence and is more complex than SPM. It consists of 36 test items.

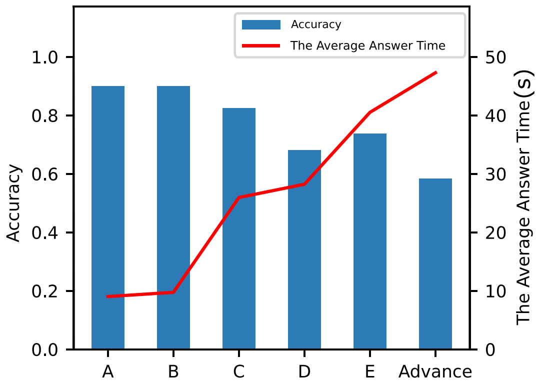


Figure 2. The accuracy and average answer time of SPM E and APM groups.

We conducted a pilot study to determine test items and time by recruiting 15 volunteer students (mean = 22.29, SD = 0.73). In this pilot study, we asked each subject to try their best to answer all the test items of APM and SPM regardless of time. We recorded the accuracy and the answering time. We produce statistics on groups (APM is considered a standalone along with five groups of SPM). As shown in figure 2, from the results of six groups (five groups of SPM plus APM as a whole group), it is obvious that participants took longer to answer the test items but performed worse as the difficulty increased. The accuracy of the SPM E group and APM group is around 60% along with the answering time of more than 40 s.

The adult with regular attention can stay focused on a task within 20 min [78]. Therefore, the entire test time is controlled within 20 min. After careful consideration, we chose the SPM E group and the APM group, containing 48 test items. The test time of each test item was set to 15 s.

3.3. Labeling technique

Most current EEG databases are labeled based on self-report data from questionnaires, such as self-assessment manikin (SAM) [79]. This method depends on personal subjective feelings. The individual difference in subjective perception may result in nuance between the label and ground truth. Therefore, this labeling technique can not ensure the accuracy of labeling. We propose a subjective and objective joint labeling technique to provide reliable labels.

For CAL, we recorded the objective performance of all participants for each test item (Correct or Incorrect). We also asked participants to complete a questionnaire to report their confusion state for each test item. As shown in table 2, confusion-related states were labeled based on objective performance and self-reported perception. If the participant answered incorrectly and expressed confusion in the questionnaire, we labeled this emotional state as 'Confused'. If the participant answered correctly and did not feel confused, we labeled the corresponding EEG as 'Non-confused'. Guessing behavior is common in examination situations, particularly when a student is unable to progress with a test item. In such cases, some students may use various tactics in an effort to arrive at the correct answer, such as looking for grammatical hints, eliminating unlikely choices, or choosing the opposite answer if other options seem similar. Others may simply choose an answer at random. These guessing behaviors can increase the chances of getting the right answer. In our experiment, we classified students' responses as 'Guess' if the student answered a test item correctly but showed confusion on the self-report questionnaire, and as 'Think-right' if the student answered a test item incorrectly but showed no confusion. 'Think-right' means that the participants thought they were right, but they were not. Overall, the aim of our labeling technique is to ensure that the state is labeled as accurately as possible.

Table 2. Four categories of states.

	Self-assessment	
Performance	Did not feel confused	Felt confused
Correct	Non-confused	Guess
Incorrect	Think-right	Confused

3.4. Smooth interaction

EEG signals are weak and easily contaminated by eye and body movement noises. Unlike other databases adopting video-based stimuli experiments, our experiment inevitably requires interaction with the computer during answering. We employ two ways to ensure the quality of the collected EEG data. One is to reduce unnecessary interactions. The user interface that we proposed only requires participants to respond with a few keys on the number keypad instead of moving the mouse to ensure that the EEG data are collected with the least noise. Thus, participants only needed to move their fingers to answer the test items. The other is to reduce unnecessary physical movement. The participants are required to keep their arms and main body still during experiments.

4. Experiment setup

4.1. Configuration and EEG recordings

The experiment system comprised three parts: EEG collector, confusion inducer, and data storage, as shown in figure 3. We employed OpenBCI as the EEG collector in this work, due to its wearable and high-quality bio-sensing hardware for brain-computer interfacing. It has eight channels (Fp1, Fp2, C3, C4, T5, T6, O1, O2) and a good sampling rate (250 Hz), providing the possibility of large-scale collection of EEG signals in learning study and application. We

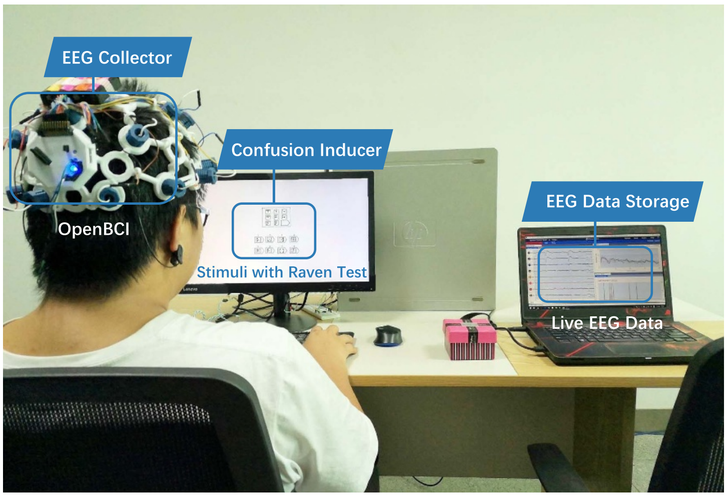


Figure 3. The experiment setup.

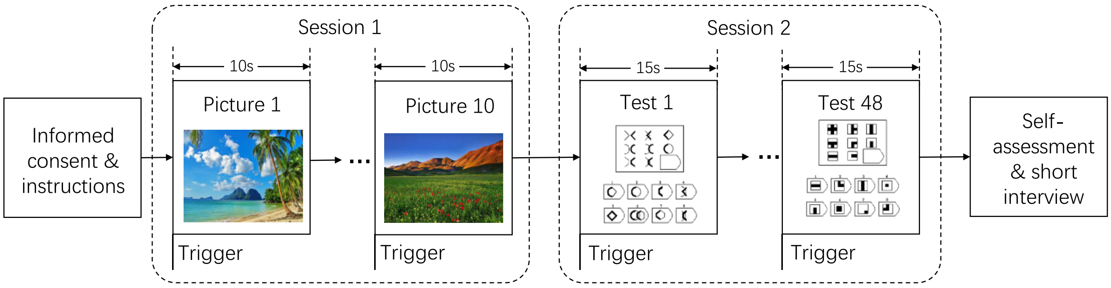


Figure 4. The experiment procedure.

used one desktop to induce confused states and a laptop to connect with the EEG collector and store the data. The E-prime [80], a software for behavioral and psychological research, was employed to generate the stimuli and interaction. It also sent trigger signals for segmenting trials. We redeveloped the firmware and software of the OpenBCI Cyton Board, making it receive the trigger signals from DB25. When storing the data, the EEG waves were real-time sync visualized. This visualization could help testers monitor the experiment.

4.2. Subjects and experiment protocol

A total of 25 subjects participated in this experiment. We obtained 23 subjects' data because the unexpected equipment problem made failed collection for two persons. The ages were between 20 and 47 (mean = 24.48, SD = 6.36); the male to female ratio is roughly half to half (12:11). The education backgrounds covered middle school, undergraduate, master, and doctoral degrees, and the major included computer science, microelectronics, bioengineering, and British and American literature. All participants were in good health and had normal vision without any history of brain injury or mental illness.

The experiment testers explained the experiment's purpose, process, and precautions. After signed an ultimate consent form, subjects started to perform the tasks. As shown in figure 4, we first presented the manipulation instruction. When subjects were ready, they watched ten scene pictures, each of which lasted 10 s. Next to this, they viewed and performed 48 tests, each of which lasted a maximum of 15 s. The participants evaluated their own level of confusion for each test item at the end of the trials.

5. Confused emotion analysis

In this section, we first evaluate the induction and labeling of confusion and then analyze band power differences between confused and non-confused states.

5.1. Confusion elicitation analysis and labeling verification

Before performing further study, it is vital to determine whether the collected data satisfied the requirements that the subject gets confused when performing some tasks and whether the EEG data are labeled correctly. Thus, we analyze the data to investigate the following two aspects:

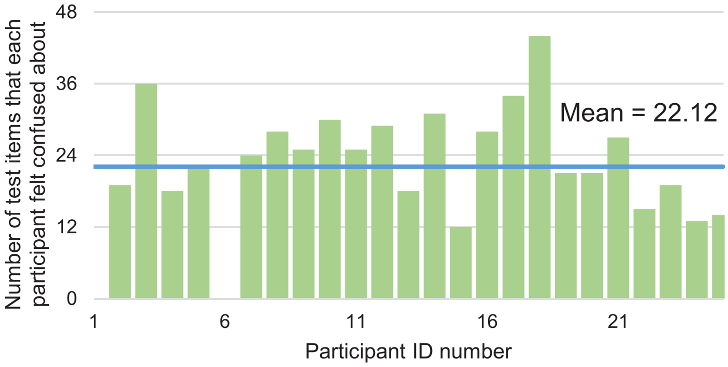


Figure 5. The number of test items that each participant felt confused about.

5.1.1. Confusion elicitation analysis

To know whether the elicitation met the requirements, we analyzed the number of test items that each participant felt confused about.

We calculated the number of test items that each participant felt confused about based on the data from their self-reported levels of confusion. In figure 5, the distribution of the self-assessment of confusion of all participants can be observed. The data of participant No. 1 and 6 were removed as we mentioned in section 4.2. Participants felt confused about at least 12 test items (i.e. participant No. 15) and at most 44 test items (i.e. participant No. 18). On average, each participant was confused by 22 test items (mean = 22.12). We also examined and counted the number of test items on which approximately half of the subjects (i.e. 12 participants) were confused. As shown in figure 5, we found that more than half of the participants were confused by 24 test items.

Overall, our induction avoids the possibility that the stimuli were designed to be challenging and confusing, but participants did not perceive them as such. The findings of the number of test items that each participant felt confused about demonstrated that confusion had been successfully triggered.

5.1.2. Labeling verification

To verify the effectiveness of proposed labeling technique, we analyzed the distribution of induced states, and compared confused and incorrect induced by each test.

First, we analyzed the distribution of induced states. As illustrated in figure 6, the pie chart depicts five states associated with the confusion emotion in the experiment. The definitions of four states, i.e. confused, non-confused, guess and think-right, are clarified in section 3.3. The term 'undefined' refers to the participants' error operations. Suppose that the participants are faced with two consecutive test items A and B. However, they ran out of time while answering test item A. As a result, the eliciting procedure automatically moved on to test item B. The participants managed to answer test item B but not test item A. This led to incorrect answers for both test items. The response for test item A was incorrect because the participants did not finish it in time, and the response for test item B was incorrect due to a misclick. The time taken to answer test item B was less than one second, which suggests that the answer was not a result of thoughtful consideration but rather a participant's response lag. As a result, the EEG data for test item B is considered undefined.

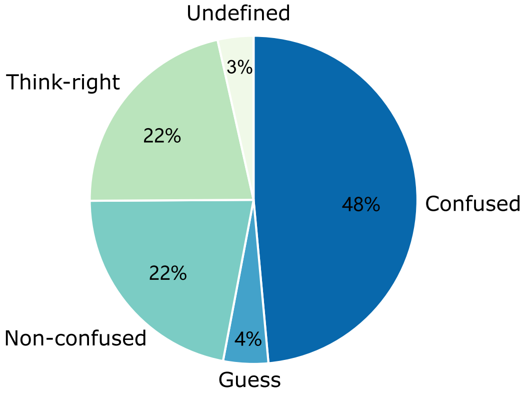


Figure 6. The distribution of induced states.

The confused states account for approximately half of all test data (48%), reaching our goal of selecting more challenging test items to trigger the confused states of participants. The following smaller proportions are non-confused and think-right states, both at 22%. Participants subjectively conveyed their non-confusion in these two situations, but in the first, they were aware that the issue had been resolved, and in the second, they believed it had. The 'think-right' state accounts for a relatively high proportion, demonstrating that people's tendency to make mistakes and have excessive self-confidence are not uncommon in the reasoning test. Since the individuals were instructed to concentrate on answering the test items

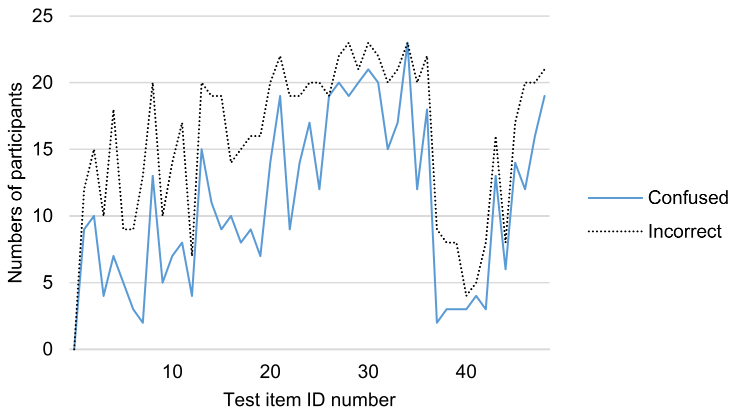


Figure 7. The comparison of 'confused' and 'incorrect' induced by each test.

and to keep trying to solve the problems, there was very little guessing. The percentage of data categorized as 'undefined' is likewise extremely low.

Second, we compared the number of participants who gave wrong responses for each test item and the number of confused participants. As shown in figure 7, for each test, the number of participants who answered wrongly were not the same as the number of participants who struggled to find a solution and felt confused. This indicated that the data labeled as 'incorrect' nuanced with the data categorized as 'confused'.

Therefore, the distribution of induced states and comparison of confused and incorrect induced by each test demonstrated that our proposed subjective and objective joint labeling technique produced labels with a high degree of granularity and reliability.

Overall, the outcomes confirmed our expectations on triggering confusion and satisfy the level of difficulty in section 3.2. This demonstrates the success of our experiment, and the EEG data can be applied to additional analysis and recognition tasks.

5.2. Band power differences between confused and non-confused states

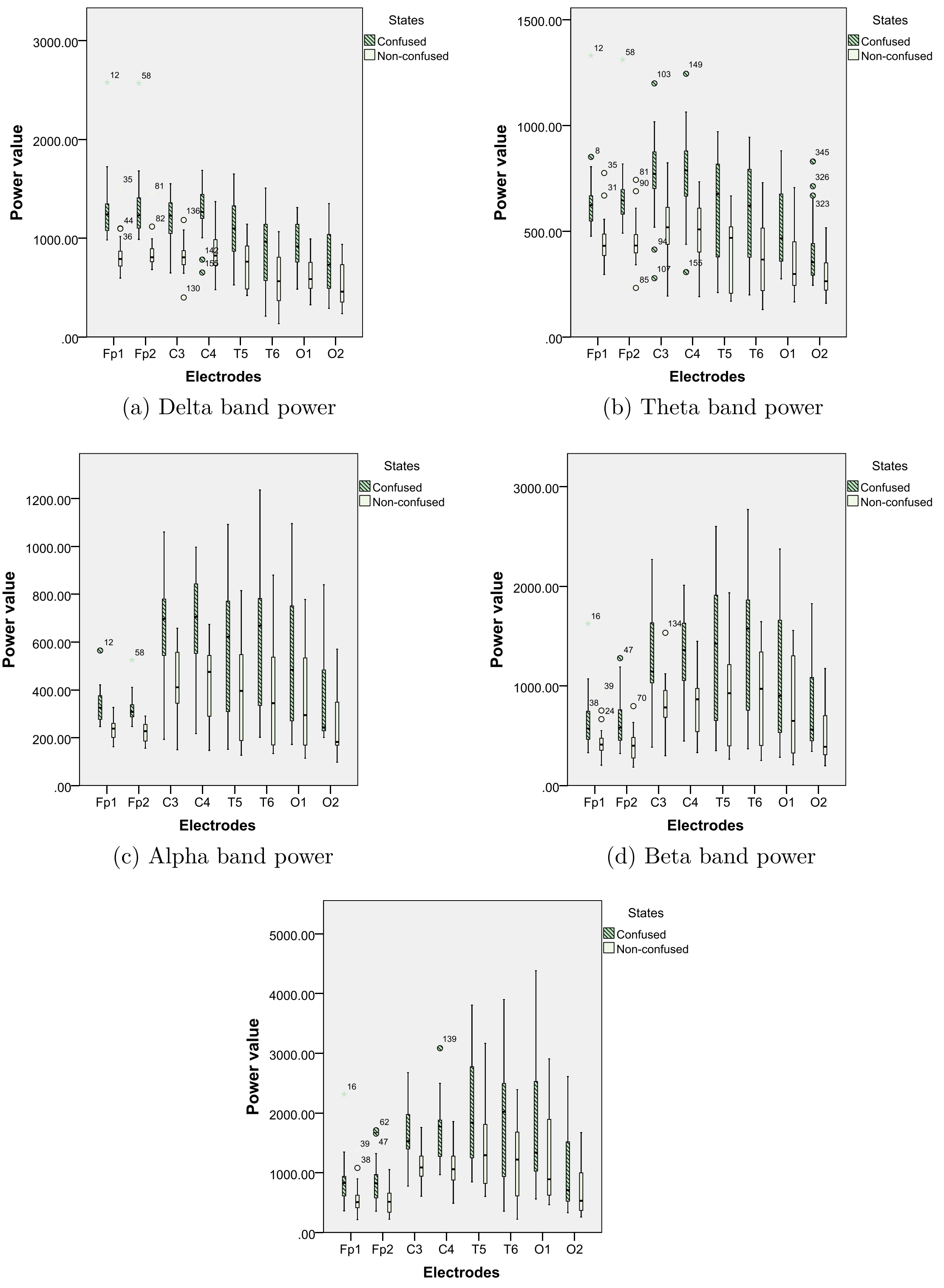
Brain activity distinctions between confused and non-confused states are a precondition for recognizing confusion in learning. However, little research has been conducted on the differences in band power between confused and non-confused states. To fill this gap, we hypothesize that there are significant differences between confused and non-confused states and assess the power of five typical bands. We calculated the band power of delta wave, theta wave, alpha wave, beta wave, and lower gamma wave for each trial. For band power, we filter each channel of the EEG data in five frequency bands: delta wave: 2–4 Hz, theta wave: 4–8 Hz, alpha wave: 8–12 Hz, beta wave: 12–30 Hz, and gamma wave: 30–70 Hz by fast Fourier transform filter. Power time courses were segmented in to 4 s.

To tackle individual differences, we computed and compared the mean band power of confused and non-confused states across all corresponding trials. As shown in figure 8, it is obvious that in all bands, confusing emotion has a greater power than non-confused emotion. This showed that the confused state and non-confused state are different, and can be distinguished based on EEG signals.

Kolmogorov–Smirnov test of observed values (differences between scores) and visual inspections of their histograms, normal Q–Q plots and box plots, showed that band power value data were not normally distributed. Thus, we employed the Wilcoxon signed ranks test to analyze data. The results are presented in table 3. The findings indicated that for delta, theta, alpha, beta, and lower gamma band power, the difference reached significance throughout all eight electrode locations (Fp1, Fp2, C3, C4, T5, T6, O1, and O2) in the confused condition as compared to non-confused condition.

6. The methodology of confusion recognition

We use two approaches to detect confusion: traditional machine learning and end-to-end techniques. The traditional machine learning approach consists of three steps: data pre-processing, feature extraction, and classification. The end-to-end methods work directly on the raw data. Pre-processing is performed using MNE and Scipy libraries, while feature extraction is done by us using Numpy and Scipy. The traditional machine learning algorithms include SVM [81], Random Forest [82], Bayes Network [83], and ANNs [84]. SVM, Random Forest, and Naive Bayes are implemented using the scikit-learn library. The SVM kernel function used is set to radial basis function, and the number of Random Forest trees is set to 80. The deep learning algorithms include long



(a) Delta band power

(b) Theta band power

(c) Alpha band power

(d) Beta band power

(e) Lower gamma band power

Figure 8. Band power comparison between confusion and non-confusion.

short term memory (LSTM) [85], ResNet [86], and EEGNet [87], which are implemented using the Pytorch library and operate directly on raw data.

The present work adopt subject-dependent and subject-independent approaches and focused on two tasks:

Table 3. Wilcoxon signed ranks tests for EEG band power measures across conditions.

		Confused		Non-confused				
		M	SD	M	SD	Z	r	Sig.
delta	Fp1	1292.86	332.35	840.50	197.48	−4.107	−0.821	0.000
	Fp2	1316.04	329.48	852.49	168.10	−4.107	−0.821	0.000
	C3	1198.54	220.08	803.65	158.94	−4.107	−0.821	0.000
	C4	1271.77	244.22	840.76	201.84	−4.107	−0.821	0.000
	T5	1097.21	342.27	737.53	237.39	−4.107	−0.821	0.000
	T6	898.73	385.24	592.08	269.47	−4.107	−0.821	0.000
	O1	934.34	240.55	629.04	191.40	−4.107	−0.821	0.000
	O2	761.18	315.84	521.82	217.79	−4.107	−0.821	0.000
theta	Fp1	656.06	175.24	454.46	110.32	−4.074	−0.815	0.000
	Fp2	666.92	165.46	458.88	114.48	−4.107	−0.821	0.000
	C3	762.43	200.62	513.42	134.95	−4.107	−0.821	0.000
	C4	767.71	210.65	501.63	143.56	−4.107	−0.821	0.000
	T5	615.08	239.12	411.40	165.41	−4.107	−0.821	0.000
	T6	569.99	247.67	377.08	176.93	−4.107	−0.821	0.000
	O1	520.04	202.20	352.49	152.44	−4.107	−0.821	0.000
	O2	404.83	166.03	292.37	104.47	−4.107	−0.821	0.000
alpha	Fp1	333.44	75.29	234.21	44.20	−4.107	−0.821	0.000
	Fp2	319.87	60.49	220.35	41.76	−4.107	−0.821	0.000
	C3	660.77	204.09	434.04	133.04	−4.107	−0.821	0.000
	C4	689.11	216.65	430.19	147.50	−4.107	−0.821	0.000
	T5	582.33	282.40	380.88	192.84	−4.074	−0.815	0.000
	T6	590.72	306.97	381.50	210.43	−4.107	−0.821	0.000
	O1	512.32	273.99	356.88	223.68	−4.107	−0.821	0.000
	O2	367.27	207.60	260.32	144.34	−4.107	−0.821	0.000
beta	Fp1	664.74	289.29	437.96	169.96	−4.107	−0.821	0.000
	Fp2	642.55	255.85	400.16	151.49	−4.107	−0.821	0.000
	C3	1274.75	438.35	822.17	261.35	−4.107	−0.821	0.000
	C4	1338.24	415.18	829.94	279.31	−4.107	−0.821	0.000
	T5	1381.67	710.47	910.12	492.79	−4.107	−0.821	0.000
	T6	1357.54	699.31	892.86	492.60	−4.107	−0.821	0.000
	O1	1118.06	647.44	767.49	490.86	−4.107	−0.821	0.000
	O2	764.14	448.02	536.27	305.04	−4.107	−0.821	0.000
lower gamma	Fp1	854.46	416.42	572.76	271.33	−4.107	−0.821	0.000
	Fp2	846.25	350.75	527.43	222.50	−4.107	−0.821	0.000
	C3	1668.81	482.33	1107.14	273.47	−4.107	−0.821	0.000
	C4	1692.51	513.89	1059.51	309.71	−4.107	−0.821	0.000
	T5	2061.67	910.08	1389.23	672.27	−4.107	−0.821	0.000
	T6	1800.05	967.22	1200.08	685.53	−4.107	−0.821	0.000
	O1	1760.00	972.98	1230.77	730.29	−4.107	−0.821	0.000
	O2	1031.92	676.90	726.16	468.72	−4.074	−0.815	0.000

(i) **Binary classification:** confused and non-confused;

(ii) **Four-class classification:** confused, non-confused, think-right, and guess.

6.1. Pre-processing

Pre-processing is applied for the raw EEG before analysis since many noises, such as eye movements, blinks, muscle, heart, and line noise, can lead to severe problems for EEG interpretation and analysis. First, a bandpass filter between 1 Hz–70 Hz was applied to filter noise and artifacts. Second, a Notch filter was applied to remove 50 Hz harmonics caused by line noise or interference. Third, Independent Component Analysis was applied to remove eye movement and blink artifacts from EEG. Finally, the EEG data was divided into 4 s sliding windows in the light of literature [88–92]. The distribution of different categories is very different in CAL; for example, guess only accounts for only 4%, while confused accounts for almost half (48%). It will seriously affect the classification results. To tackle the issue of data imbalance, we set different length overlapping parts: 0.25 s for confused, 0.75 s for non-confused, 0.5 s for think-right, and 0.75 s for guess.

6.2. Feature extraction

We extracted six features from three categories: time domain, frequency domain and spatial domain.

For temporal features, we adopt Hjorth parameters, including activity, mobility and complexity, and the time-domain energy [93]. Because these temporal feature data dimension are small, we put them

Table 4. Binary classification results (Acc/F1 score).

Subject-dependent approach				
Feature	Naive Bayes	SVM	Random Forest	ANN
Temporal mixed feature	55.49/0.53	50.71/0.33	**88.06/0.88**	66.00/0.65
DE	50.71/0.33	50.71/0.33	50.71/0.33	58.64/0.56
PSD	57.30/0.52	67.43/0.67	69.72/0.69	69.63/0.69
Band power	53.39/0.51	58.73/0.54	77.17/0.76	72.97/0.72
DASM	51.76/0.39	61.79/0.60	60.64/0.59	59.22/0.58
RASM	51.19/0.35	51.19/0.35	60.84/0.59	58.74/0.57

Table 5. Four classification results (Acc/F1 score).

Subject-dependent approach				
Feature	Naive Bayes	SVM	Random Forest	ANN
Temporal mixed feature	38.35/0.29	34.24/0.13	**73.41/0.72**	41.90/0.29
DE	33.60/0.12	33.60/0.12	47.34/0.38	33.61/ 0.12
PSD	50.06/0.39	48.10/0.33	37.72/0.26	50.13/0.43
Band power	61.45/0.53	38.16/0.21	36.13/0.25	51.14/0.40
DASM	41.32/0.29	41.64/0.27	34.36/0.15	38.99/0.27
RASM	42.84/0.32	33.73/0.33	33.92/0.13	38.29/0.25

together as a mixed feature, called temporal mixed feature.

For frequency features, we select three types of features: differential entropy (DE) [94], power spectral density (PSD) [95] and band power. DE is defined as follows:

$$h_i(X) = \frac{1}{2}\log(2\pi e\sigma_i^2) \tag{1}$$

where h_i and σ_i^2 denote the DE of the corresponding EEG signal in frequency band i and the signal variance, respectively

PSD is computed by the Welch method in this work, defined as follows:

$$P(f) = \frac{1}{MU}\left|\sum_{n=0}^{M-1} x_i(n)w(n)e^{-j2\pi f}\right|^2 \tag{2}$$

$$P_{\text{Welch}}(f) = \frac{1}{L}\sum_{i=0}^{L-1} P(f) \tag{3}$$

where M is interval length, U stands for normalization factor for power in window function.

The details of band power calculation are described in section 5.2.

For spatial features, we choose the two features: differential asymmetry (DASM) and rational asymmetry (RASM) [94]. DASM and RASM refers the differences and ratios between DE, defined as follows:

$$\text{DASM} = \text{DE}(X_i^{\text{left}}) - \text{DE}(X_i^{\text{right}}) \tag{4}$$

$$\text{RASM} = \text{DE}(X_i^{\text{left}})/\text{DE}(X_i^{\text{right}}) \tag{5}$$

where X_i^{left} and X_i^{right} refers to electrode pairs. This work used Fp1-Fp2, C3-C4, T5-T6, and O1-O2 as pairs.

6.3. Subjects-dependent/independent approaches

Towards emotion recognition algorithm, there are two types of approaches: subject-dependent and subject-independent. In the subject-dependent model, the classifier is trained for each subject individually, while in the subject-independent model, the classifier is trained across multiple subjects. To fully evaluate CAL, we employ conventional machine learning and deep learning to explore binary classification tasks (confused, non-confused) and four classification tasks (confused, non-confused, think-right, and guess) on subjects-dependent and subject-independent approaches, respectively.

6.3.1. Subject-dependent results

In the subject-dependent approach, the EEG data of all trials for each subject were divided 70% for the train set and 30% for the test set. We adopt four mainstream classifiers for the classic machine learning methods: SVM, Random Forest, Bayes Network, and ANN, with features extracting spatial, frequency, and temporal domains as mentioned above.

The accuracy results on binary-class were presented in table 4. Random Forest with temporal mixed feature performs better than other machine learning algorithms, achieving 88.06% in accuracy and 0.88 in F1 score.

The performance of Random Forest with the time-domain feature (73.41%/0.72) is also better than other algorithms in four categories, as shown in table 5.

For end-to-end methods, EEGNet, ResNet, and LSTM were chosen to apply to CAL. Table 6 compares the results obtained from the binary and four classification tasks. What stands out in the table is that the ResNet achieved relatively higher accuracy (80.61%/0.80) in binary and four-class tasks.

Table 6. End-to-end methods results (Acc/F1 score).

Subject-dependent approach		
Method	Binary classification	Four classification
LSTM	73.45/0.73	53.29/0.49
ResNet	80.61/0.80	73.10/0.73
EEGNet	72.02/0.71	49.81/0.43

Table 7. Binary classification results (Acc/F1 score).

Subject-independent approach				
Feature	Naive Bayes	SVM	Random Forest	ANN
Temporal mixed feature	52.62/0.51	52.53/0.42	84.43/0.84	55.18/0.52
DE	54.10/0.35	54.19/0.36	57.76/0.57	55.71/0.50
PSD	60.59/0.55	55.71/0.53	55.98/0.52	57.59/0.57
Band power	57.05/0.56	41.78/0.41	54.82/0.52	60.45/0.58
DASM	43.48/0.35	50.35/0.48	48.57/0.46	52.05/0.49
RASM	54.10/0.35	54.01/0.37	45.89/0.45	55.27/0.54

Table 8. Four classification results (Acc/F1 score).

Subject-independent approach				
Feature	Naive Bayes	SVM	Random Forest	ANN
Temporal mixed feature	34.31/0.20	36.79/0.14	36.85/0.26	40.89/0.22
DE	36.55/0.13	36.61/0.14	34.62/0.22	40.23/0.24
PSD	40.83/0.27	38.05/0.31	35.94/0.25	40.17/0.27
Band power	37.81/0.26	27.08/0.16	38.29/0.23	41.62/0.29
DASM	30.45/0.19	34.13/0.19	29.37/0.14	38.12/0.20
RASM	30.09/0.19	36.12/0.13	36.55/0.13	40.53/0.25

Overall, for the subject-dependent approach, Random Forest with the time-domain feature did the best work in both the binary classification task with 88.06%/0.88 and the four classification tasks with 73.41%/0.72.

6.3.2. Subject-independent results

In the subject-independent approach, the data was divided into 70%/30% cross subjects. The data from 16 subjects were considered the training set, while the remaining data from 7 subjects were used to test. All the experiments setting were the same as the experiment setting of the subject-dependent approach.

For the binary classification tasks, the results from four conventional machine learning methods are shown in table 7. The accuracy of Random Forest with the time-domain feature achieved 84.43%/0.84, performing better than other machine learning algorithms.

For four-class tasks, we found that the result of ANN with band power achieved the best performance (41.62%/0.29), as shown in table 8.

Table 9 presents the results of the end-to-end methods. EEGnet and LSTM provided higher accuracy than other end-to-end methods in the binary classification task (64.46%/0.61) and four classification tasks (40.53%/0.24), respectively.

Table 9. End-to-end methods results (Acc/F1 score).

Subject-independent approach		
Method	Binary classification	Four classification
LSTM	61.25/0.59	**40.53/0.24**
ResNet	57.95/0.55	40.05/0.23
EEGNet	**64.46/0.61**	39.14/0.20

Overall, we achieved 84.43%/0.84 (Random Forest with the time-domain feature) for the subject-independent approach in the binary classification and 41.62%/0.29 (ANN with band power) in four-class tasks.

7. Discussion, limitations and future directions

In this section, we discussed with answering three research questions (RQs):.

- RQ1: How can the confusion be accurately triggered and labeled in reasoning tasks?
- RQ2: What does the EEG indicate when a learner becomes confused?
- RQ3: What is the best way to recognize confusion in the given situation?

Furthermore, limitations and future directions are also considered.

7.1. Discussion

7.1.1. How can the confusion be accurately triggered and labeled in reasoning tasks?

The theories that emphasize the importance of cognitive disequilibrium and confusion in learning served as the foundation for the present study. The goal was to use EEG to recognize and examine learners' confused states during pattern reasoning when they encounter cognitive disequilibrium and struggle to come up with a solution. Since this goal required a scenario that could trigger participants to get confused, we employed Raven's tests to build pattern reasoning tasks. The confusion induction was one focus of the first research question (RQ1). By examining how participants felt in the experiment, we found that tests made participants engage well, and the confusion had been successfully triggered. Tests were therefore shown to be active stimuli to invoke confusion. By varying the difficulty of test items and imposing a time limit on responses, researchers can trigger confusion as expected. Furthermore, the results showed that confusion occurs outside of knowledge absorption, which is sometimes referred to as knowledge input. It also occurs in the output stages, like in tasks requiring reasoning. Existing theories of confusion and cognitive disequilibrium were built upon scenarios of information inconsistency and problem-solving. The present study refines existing theories from the perspective of reasoning, particularly rule learning.

In this study, we did not simply assign two labels to the data using self-assessment: confused and non-confused. Instead, we proposed and employed a subjective and objective joint labeling technique and found five states related to confusion: confused, non-confused, think-right, guess and undefined. The 'think-right' state refers to a state that learners believed they were correct and but they answered incorrectly. By analyzing the distribution of induced states, we found that the 'think-right' states account for a high proportion. This finding was consistent with overconfidence is common among learners [96, 97]. Eva *et al* [97] found poor correlations between medical students' estimated knowledge and actual test scores. They concluded that self-assessment is a poor predictor of actual performance. This phenomenon emphasizes the significance of precise labeling. The students overestimate their abilities and performance, so they are not psychologically confused. However, not being confused should correspond to knowledge mastery and good performance in reasoning tasks. When the data is labeled as 'non-confused' rather than 'think-right', the detection results are misleading, which make researchers, educators and practitioners think the learner was not impeded by the current materials. As a result, our labeling method improves the label's accuracy for each state associated with confusion.

7.1.2. What does the EEG indicate when a learner becomes confused?

EEG has been used to detect learning confusion, but there are no definite signs that confused states differ from non-confused states. Through analyzing band power of delta, theta, alpha, beta and lower gamma, we found that the confused state shows a greater power than non-confused one. And there were significant differences between these two states on prefrontal, central, temporal and occipital locations. The findings clearly indicated that confused states can be distinguished from non-confused states in brain activity.

As a result of functional modulation during cognitive tasks, changes in brain oscillations take place in various EEG frequency bands. Theta activity is one of these task-related oscillations that has been linked to memory performance. Several studies have found that theta power increases as working memory load increases [98]. We observed that theta power was significantly higher in confused states compared with non-confused states on all eight electrodes. First, this indicated the confused state is more loaded than the non-confused, being in line with intuition. Second, this outcome was consistent with previous research that indicated a negative correlation between theta power and performance, with higher theta band power indicating lower performance [35]. Confused test takers did not provide the right response to the test item. Participants who were not confused provided accurate answers to the test items.

The attention-demanding to target stimuli can be represented by several frequency oscillations. We found that the alpha band power was significantly higher in the confused state than in the non-confused state, similar to the effects of the theta oscillation. The previously observed model only shows engagement as equilibrium and confusion as disequilibrium [1]. It was unclear whether there were differences in engagement and cognitive load between confused and non-confused states. The study [39] shows that delta oscillations are associated with cognitive functions, including decision-making and attentional activities. We observed that the power of the delta wave for the confusion state is higher than for the non-confusion state at all eight electrodes. This is consistent with the finding of previous work [38, 99] that delta energy increases during mental computation. According to our findings on delta, theta and alpha, learning targets requires more attention and cognitive resources when an individual is confused.

Confusion is not an isolated learning emotion. It changes as other cognitive or affective states shift. Existing models on cognitive-affective process related to confusion do not reveal how the learners' attention and cognitive load vary when they feel confused or not. The present work fills this gap by exploring preliminarily the relationship of attention, cognitive load and confusion. Because the difficult level of tests used in this experiment were not progressively varied, we did not compare attention, cognitive load and confusion over time. To refine this cognitive-affective process related to confusion, we plan to investigate the confusion using the stimuli tests with progressively increasing difficulty in the future work.

7.1.3. What is the best way to recognize confusion in the given situation?

Since Picard [100] proposed 'affective computing' two decades ago, researchers have attempted to detect learning emotion in digital learning environments in order to understand the learners' states and deliver appropriate feedback based on such emotions [16, 101, 102]. These intelligent emotional-adaptive systems can increase not only performance but also the learning experience. Maintaining appropriate confusion, which represents cognitive disequilibrium, helps students learn more deeply and keeps them interested in learning at all times. Therefore, it is crucial to make confusion detection feasible and practical. Our findings and CAL provide practical implications for recognizing and analyzing confusion, as well as emotional-adaptive system. First, our CAL database can assist and encourage other researchers to evaluate their proposed methods or tools for recognizing students' confused states. Second, other researchers might compare confusion in various contexts using the EEG data from CAL, to develop or broaden the confusion theory. Third, our experiment showed how different features (time domain, frequency domain, and spatial domain) impacted the recognition results. To better understand the cognitive process and improve accuracy, it is important to find stable features, which is a focus for future research.

7.2. Limitations and future directions

The present study has several limitations that should be addressed in future investigations. First, we used a portable device to gather and detect EEG data because this study was for educational purposes. The portable device is based on the dry electrode system and is comfortable for users to wear for extended periods. Therefore, this type of apparatus is appropriate for research and practice in educational settings. However, because of the limited electrodes, it is impossible to examine the topography and networks of the brain using portable equipment. Hence, it is interesting to investigate these differences between confused and non-confused states using gel-based or sponge-based devices with 32 or 64 electrodes. This will contribute to a full comprehension of the characteristics of confused emotions.

Second, self-reports were obtained after performing the full test items. Because the experiment was short, participants had no difficulty recalling whether they were confused or not about each item. However, if the experiment lasts a long time, the validity is dependent on the participants' memory, which is outside the researcher's control. One option is to question the participant after each trial/test item. But this may cause problems, such as influencing learners' thinking and disrupting their flow. There is a compromise between extending experiment time and maintaining recall validity. Thus, pilot studies should be conducted to strike a balance.

The third limitation concerns the sensitivity of our labelling technique in eliminating other emotions. This technique allows states to be identified more accurately and therefore more states are labeled. However, the active emotion elicitation method we used in this study is naturalistic and similar to a real emotional event. It is generally difficult to manipulate accurately and results in a wider range of emotional responses from individuals [29]. Therefore, confusion may not be a single emotion during the experiment. To address this issue, we plan to improve the subjective scale in our future work and propose a novel scale to capture all emotions experienced by subjects during the task, as well as the arousal and dominance of these emotions.

Finally, we did not collect eye-tracking data. Eye-tracking technology supports the direct and objective recording of learners' eye movements in real-time, allowing visual attentional processes to be observed, quantified, and analyzed [103–107]. Previous educational psychology studies commonly used eye-trackers to explore learners' attention, such as using an eye-tracker to analyze instructor presence in video lectures [108]. By eye-tracking, it is possible to figure out which part of the materials or interface causes the learners to feel confused. As a result, we plan to collect data from multiple sources (EEG, eye-tracking, and video) in our near future work.

8. Conclusion

Decades of research have proven the impact of confusion on learning, but little is known about its associated brain activities, the cognitive-affective intertwining processes, and also detection in reasoning. The present research empirically contributed to this area by investigating the confused states during reasoning, establishing that confused and non-confused states can be classified, exploring the relationship between attention, cognitive load and confusion at the level of brain, and providing an EEG database and benchmark results for recognizing confusion. It also contributed theoretically by expanding existing theories on the differences between confused and non-confused states when people learn, as well as the cognitive-affective model involving attention, cognitive load and confusion. The practical contribution of this work allows researchers, educators and practitioners to monitor confusion in their teaching environments, build their adaptive system, and test their proposed detection approaches.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: https://intelligentinteractivelab.github.io/datasets/cal.

Acknowledgments

This work was supported by the National Natural Science Foundation of China: 62077036; National Key Research and Development Program of China: 2022YFC3303600, 2018AAA0100501; We gratefully acknowledge the support of Ascend CANN (Compute Architecture for Neural Networks) and Ascend 910 AI Processor used for this research. Artificial intelligence computing power and technical support are provided by Xi'an Future Artificial Intelligence Computing Center. The author's have confirmed that any identifiable participants in this study have given their consent for publication.

Ethical statement

This research was approved by the Ethics Committee of the Faculty of Education at Shaanxi Normal University. The participants were volunteers who provided written informed consent. They were informed that they had the right to withdraw from the study at any time without penalty. Confidentiality was ensured by using numbers instead of names in the research database. Data were only used for research purposes. This study was performed in accordance with the 1964 declaration of HELSINKI and later amendments.

ORCID iDs

Tao Xu https://orcid.org/0000-0002-1721-561X
Jiabao Wang https://orcid.org/0000-0002-8889-3223
Gaotian Zhang https://orcid.org/0000-0001-5272-5298
Ling Zhang https://orcid.org/0000-0001-5174-4438
Yun Zhou https://orcid.org/0000-0002-2306-8986

References

[1] D'Mello S and Graesser A 2012 Dynamics of affective states during complex learning *Learn. Instr.* **22** 145–57

[2] Graesser A C and D'Mello S 2012 Emotions during the learning of difficult material *The Psychology of Learning and Motivation* (*Psychology of Learning and Motivation* vol 57) ed B H Ross (New York: Academic) ch 5, pp 183–225

[3] D'Mello S, Dale R and Graesser A 2012 Disequilibrium in the mind, disharmony in the body *Cogn. Emot.* **26** 362–74

[4] Graesser A C, Lu S, Olde B A, Cooper-Pye E and Whitten S 2005 Question asking and eye tracking during cognitive disequilibrium: comprehending illustrated texts on devices when the devices break down *Mem. Cogn.* **33** 1235–47

[5] Baker R S J D, D'Mello S K, Rodrigo M M T and Graesser A C 2010 Better to be frustrated than bored: the incidence, persistence and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments *Int. J. Hum.-Comput. Stud.* **68** 223–41

[6] Zhou Y, Xu T, Li S and Li S 2018 Confusion state induction and EEG-based detection in learning *2018 40th Annual Int. Conf. IEEE Engineering in Medicine and Biology Society (EMBC)* pp 3290–3

[7] Han Z-M, Huang C-Q, Yu J-H and Tsai C-C 2021 Identifying patterns of epistemic emotions with respect to interactions in massive online open courses using deep learning and social network analysis *Comput. Hum. Behav.* **122** 106843

[8] D'Mello S and Graesser A 2014 Confusion and its dynamics during device comprehension with breakdown scenarios *Acta Psychol.* **151** 106–16

[9] Arguel A, Lockyer L, Kennedy G, Lodge J M and Pachman M 2019 Seeking optimal confusion: a review on epistemic emotion management in interactive digital learning environments *Interact. Learn. Environ.* **27** 200–10

[10] Zhou Y, Xu T, Li S and Shi R 2019 Beyond engagement: an EEG-based methodology for assessing user's confusion in an educational game *Univers. Access Inf. Soc.* **18** 551–63

[11] Lehman B, Matthews M, D'Mello S and Person N 2008 What are you feeling? Investigating student affective states during expert human tutoring sessions *Intelligent Tutoring Systems* ed B P Woolf, E Aïmeur, R Nkambou and S Lajoie (Berlin: Springer) pp 50–59

[12] Lehman B, D'Mello S and Graesser A 2012 Confusion and complex learning during interactions with computer learning environments *Internet Higher Educ.* **15** 184–94

[13] D'Mello S, Lehman B, Pekrun R and Graesser A 2014 Confusion can be beneficial for learning *Learn. Instr.* **29** 153–70

[14] Vogl E, Pekrun R, Murayama K, Loderer K and Schubert S 2019 Surprise, curiosity and confusion promote knowledge exploration: evidence for robust effects of epistemic emotions *Frontiers Psychol.* **10** 2474

[15] Arguel A, Lockyer L, Chai K, Pachman M and Lipp O V 2019 Puzzle-solving activity as an indicator of epistemic confusion *Frontiers Psychol.* **10** 163

[16] Malekzadeh M, Mustafa M B and Lahsasna A 2015 A review of emotion regulation in intelligent tutoring systems *J. Educ. Technol. Soc.* **18** 435–45

[17] Sullins J and Graesser A C 2014 The relationship between cognitive disequilibrium, emotions and individual differences on student question generation *Int. J. Learn. Technol.* **9** 221–47

[18] D'Mello S K, Craig S D, Witherspoon A, McDaniel B and Graesser A 2008 Automatic detection of learner's affect from conversational cues *User Model. User-Adapt. Interact.* **18** 45–80

[19] Pachman M, Arguel A, Lockyer L, Kennedy G and Lodge J 2016 Eye tracking and early detection of confusion in digital learning environments: proof of concept *Australas. J. Educ. Technol.* **32** 58–71

[20] Wang H, Li Y, Hu X, Yang Y, Meng Z and Chang K-M 2013 Using EEG to improve massive open online courses feedback interaction *Proc. 1st Workshop on Massive Open Online Courses at the 16th Annual Conf. on Artificial Intelligence in Education* vol 1009 pp 59–66

[21] Berka C *et al* 2007 EEG correlates of task engagement and mental workload in vigilance, learning and memory tasks *Aviat. Space Environ. Med.* **78** B231–44

[22] Xu T, Zhou Y, Wang Z and Peng Y 2018 Learning emotions EEG-based recognition and brain activity: a survey study on BCI for intelligent tutoring system *Proc. Comput. Sci.* **130** 376–82

[23] Galán F C and Beal C R 2012 EEG estimates of engagement and cognitive workload predict math problem solving outcomes *Int. Conf. on User Modeling, Adaptation and Personalization* (Springer) pp 51–62

[24] Vasiljevic G A M and de Miranda L C 2020 Brain–computer interface games based on consumer-grade EEG devices: a systematic literature review *Int. J. Hum.-Comput. Interact.* **36** 105–42

[25] Lin F-R and Kao C-M 2018 Mental effort detection using EEG data in E-learning contexts *Comput. Educ.* **122** 63–79

[26] Xu J and Zhong B 2018 Review on portable EEG technology in educational research *Comput. Hum. Behav.* **81** 340–9

[27] Liang Y, Liu X, Qiu L and Zhang S 2018 An EEG study of a confusing state induced by information insufficiency during mathematical problem-solving and reasoning *Comput. Intell. Neurosci.* **2018** 1943565

[28] Reñosa C R M, Bandala A A and Vicerra R R P 2019 Classification of confusion level using EEG data and artificial neural networks *2019 IEEE 11th Int. Conf. on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM)* pp 1–6

[29] Hu X, Chen J, Wang F and Zhang D 2019 Ten challenges for EEG-based affective computing *Brain Sci. Adv.* **5** 1–20

[30] Seel N M 2012 Inferential learning and reasoning *Encyclopedia of the Sciences of Learning* ed N M Seel (Boston, MA: Springer) pp 1550–5

[31] Pekrun R and Linnenbrink-Garcia L 2012 Academic emotions and student engagement *Handbook of Research on Student Engagement* (Boston, MA: Springer) pp 259–82

[32] Lodge J M, Kennedy G, Lockyer L, Arguel A and Pachman M 2018 Understanding difficulties and resulting confusion in learning: an integrative review *Front. Educ.* **3** 49

[33] Rob B K, Reilly R and Picard R W 2001 External representation of learning process and domain knowledge: affective state as a determinate of its structure and function *Artificial Intelligence in Education Workshops* pp 64–69

[34] Cohen M X 2014 *Analyzing Neural Time Series Data: Theory and Practice* (*Issues in Clinical and Cognitive Neuropsychology*) (Cambridge, MA: MIT Press)

[35] Klimesch W 1999 EEG alpha and theta oscillations reflect cognitive and memory performance: a review and analysis *Brain Res. Rev.* **29** 169–95

[36] Başar E, Başar-Eroğlu C, Karakaş S and Schürmann M 1999 Are cognitive processes manifested in event-related gamma, alpha, theta and delta oscillations in the EEG? *Neurosci. Lett.* **259** 165–8

[37] Başar E, Başar-Eroğlu C, Karakaş S and Schürmann M 2000 Brain oscillations in perception and memory *Int. J. Psychophysiol.* **35** 95–124

[38] Harmony T 2013 The functional significance of delta oscillations in cognitive processing *Front. Integr. Neurosci.* **7** 83

[39] Güntekin B and Başar E 2016 Review of evoked and event-related delta responses in the human brain *Int. J. Psychophysiol.* **103** 43–52

[40] Pope A T, Bogart E H and Bartolome D S 1995 Biocybernetic system evaluates indices of operator engagement in automated task *Biol. Psychol.* **40** 187–95

[41] Szafir D and Mutlu B 2012 Pay attention! designing adaptive agents that monitor and improve user engagement *Proc. SIGCHI Conf. on Human Factors in Computing Systems (CHI '12)* (New York: Association for Computing Machinery) pp 11–20

[42] Andujar M and Gilbert J E 2013 Let's learn! enhancing user's engagement levels through passive brain-computer interfaces *CHI '13 Extended Abstracts on Human Factors in Computing Systems (CHI EA '13)* (New York: Association for Computing Machinery) pp 703–8

[43] Huang J, Yu C, Wang Y, Zhao Y, Liu S, Mo C, Liu J, Zhang L and Shi Y 2014 Focus: enhancing children's engagement in reading by using contextual BCI training sessions *Proc. SIGCHI Conf. on Human Factors in Computing Systems (CHI '14)* (New York: Association for Computing Machinery) pp 1905–8

[44] Xu T, Wang X, Wang J and Zhou Y 2021 From textbook to teacher: an adaptive intelligent tutoring system based on BCI *2021 43rd Annual Int. Conf. IEEE Engineering in Medicine & Biology Society (EMBC)* pp 7621–4

[45] Grimes D, Tan D S, Hudson S E, Shenoy P and Rao R P N 2008 Feasibility and pragmatics of classifying working memory load with an electroencephalograph *Proc. SIGCHI Conf. on Human Factors in Computing Systems (CHI '08)* (New York: Association for Computing Machinery) pp 835–44

[46] Gerjets P, Walter C, Rosenstiel W, Bogdan M and Zander T O 2014 Cognitive state monitoring and the design of adaptive instruction in digital environments: lessons learned from cognitive workload assessment using a passive brain–computer interface approach *Front. Neurosci.* **8** 385

[47] Gevins A, Smith M E, Leong H, McEvoy L, Whitfield S, Du R and Rush G 1998 Monitoring working memory load during computer-based tasks with EEG pattern recognition methods *Hum. Factors* **40** 79–91

[48] Jensen O and Tesche C D 2002 Frontal theta activity in humans increases with memory load in a working memory task *Eur. J. Neurosci.* **15** 1395–9

[49] Gevins A and Smith M E 2003 Neurophysiological measures of cognitive workload during human–computer interaction *Theor. Issues Ergon. Sci.* **4** 113–31

[50] Missonnier P, Deiber M-P, Gold G, Millet P, Gex-Fabry Pun M, Fazio-Costa L, Giannakopoulos P and Ibáñez V 2006 Frontal theta event-related synchronization: comparison of directed attention and working memory load effects *J. Neural Transm.* **113** 1477–86

[51] Sauseng P, Griesmayr B, Freunberger R and Klimesch W 2010 Control mechanisms in working memory: a possible function of EEG theta oscillations *Neurosci. Biobehav. Rev.* **34** 1015–22

[52] Lei S and Roetting M 2011 Influence of task combination on EEG spectrum modulation for driver workload estimation *Hum. Factors* **53** 168–79

[53] Gevins A, Smith M E, McEvoy L and Yu D 1997 High-resolution EEG mapping of cortical activation related to working memory: effects of task difficulty, type of processing and practice *Cereb. Cortex* **7** 374–85

[54] Stipacek A, Grabner R H, Neuper C, Fink A and Neubauer A C 2003 Sensitivity of human EEG alpha band desynchronization to different working memory components and increasing levels of memory load *Neurosci. Lett.* **353** 193–6

[55] Kamzanova A T, Kustubayeva A M and Matthews G 2014 Use of EEG workload indices for diagnostic monitoring of vigilance decrement *Hum. Factors* **56** 1136–49

[56] Puma S, Matton N, Paubel P-V, Raufaste E and El-Yagoubi R 2018 Using theta and alpha band power to assess cognitive workload in multitasking environments *Int. J. Psychophysiol.* **123** 111–20

[57] 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 *J. Educ. Comput. Res.* **55** 526–51

[58] Wang H, Li Y, Hu X, Yang Y, Meng Z and Chang K-M 2013 Using EEG to improve massive open online courses feedback interaction *AI-ED Workshop Proc.* pp 59–66

[59] Lotte F, Bougrain L, Cichocki A, Clerc M, Congedo M, Rakotomamonjy A and Yger F 2018 A review of classification algorithms for EEG-based brain-computer interfaces: a 10-year update *J. Neural Eng.* **15** 031005

[60] LeCun Y, Bengio Y and Hinton G 2015 Deep learning *Nature* **521** 436–44

[61] Schirrmeister R T, Springenberg J T, Fiederer L D J, Glasstetter M, Eggensperger K, Tangermann M, Hutter F, Burgard W and Ball T 2017 Deep learning with

convolutional neural networks for EEG decoding and visualization *Hum. Brain Mapp.* **38** 5391–420

[62] Liu Y-J, Yu M, Zhao G, Song J, Ge Y and Shi Y 2018 Real-time movie-induced discrete emotion recognition from EEG signals *IEEE Trans. Affect. Comput.* **9** 550–62

[63] Koelstra S, Muhl C, Soleymani M, Lee J-S, Yazdani A, Ebrahimi T, Pun T, Nijholt A and Patras I 2012 DEAP: a database for emotion analysis; using physiological signals *IEEE Trans. Affect. Comput.* **3** 18–31

[64] Soleymani M, Lichtenauer J, Pun T and Pantic M 2012 A multimodal database for affect recognition and implicit tagging *IEEE Trans. Affect. Comput.* **3** 42–55

[65] Zheng W-L and Lu B-L 2015 Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks *IEEE Trans. Auton. Mental Dev.* **7** 162–75

[66] Becker H, Fleureau J, Guillotel P, Wendling F, Merlet I and Albera L 2017 Emotion recognition based on high-resolution EEG recordings and reconstructed brain sources *IEEE Trans. Affect. Comput.* **11** 1

[67] Subramanian R, Wache J, Abadi M K, Vieriu R L, Winkler S and Sebe N 2018 ASCERTAIN: emotion and personality recognition using commercial sensors *IEEE Trans. Affect. Comput.* **9** 147–60

[68] Correa J A M, Abadi M K, Sebe N and Patras I 2018 AMIGOS: a dataset for affect, personality and mood research on individuals and groups *IEEE Trans. Affect. Comput.* **12** 1

[69] Katsigiannis S and Ramzan N 2018 DREAMER: a database for emotion recognition through EEG and ECG signals from wireless low-cost off-the-shelf devices *IEEE J. Biomed. Health Inform.* **22** 98–107

[70] Song T, Zheng W, Lu C, Zong Y, Zhang X and Cui Z 2019 MPED: a multi-modal physiological emotion database for discrete emotion recognition *IEEE Access* **7** 12177–91

[71] Liu Y and Sourina O 2013 EEG databases for emotion recognition *2013 Int. Conf. on Cyberworlds* pp 302–9

[72] Mehmood R M and Lee H J 2015 EEG based emotion recognition from human brain using Hjorth parameters and SVM *Int. J. Bio-Sci. Bio-Technol.* **7** 23–32

[73] Greco A, Faes L, Catrambone V, Barbieri R, Scilingo E P and Valenza G 2019 Lateralization of directional brain-heart information transfer during visual emotional elicitation *Am. J. Physiol. Regul. Integr. Comp. Physiol.* **317** R25–R38

[74] Mehmood R M and Lee H J 2016 Toward an analysis of emotion regulation in children using late positive potential *Annual Int. Conf. IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual Int. Conf.* vol 2016 pp 279–82

[75] Murugappan M, Alshuaib W B, Bourisly A, Sruthi S, Khairunizam W, Shalini B and Yean W 2020 Emotion classification in Parkinson's disease EEG using RQA and ELM *2020 16th IEEE Int. Coll. on Signal Processing & Its Applications (CSPA)* pp 290–5

[76] Raven J 2000 The Raven's progressive matrices: change and stability over culture and time *Cogn. Psychol.* **41** 1–48

[77] Kunda M, McGreggor K and Goel A K 2013 A computational model for solving problems from the Raven's progressive matrices intelligence test using iconic visual representations *Cogn. Syst. Res.* **22-23** 47–66

[78] Bradbury N A 2016 Attention span during lectures: 8 seconds, 10 minutes, or more? *Adv. Physiol. Educ.* **40** 509–13

[79] Morris J D 1995 Observations: SAM: the self-assessment manikin: an efficient cross-cultural measurement of emotional response *J. Advert. Res.* **35** 63–68

[80] Psychology Software Tools 2022 E-prime 2.0 (available at: https://psychology-software-tools.mybigcommerce.com/e-prime-2-0)

[81] Kanamori T, Hido S and Sugiyama M 2009 A least-squares approach to direct importance estimation *J. Mach. Learn. Res.* **10** 1391–445

[82] Breiman L 2001 Random forests *Mach. Learn.* **45** 5–32

[83] Zhang H 2004 The optimality of naive Bayes *Proc. 17th Int. Florida Artificial Intelligence Research Society Conf. (Miami Beach, Florida, USA)* ed V Barr and Z Markov (AAAI Press) pp 562–7

[84] Wang S-C 2003 Artificial neural network *Interdisciplinary Computing in Java Programming* ed S-C Wang (Boston, MA: Springer) pp 81–100

[85] Sak Hşim, Senior A and Beaufays Fçoise 2014 Long short-term memory based recurrent neural network architectures for large vocabulary speech recognition (arXiv:1402.1128 [cs, stat])

[86] He K, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition *2016 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (IEEE Computer Society) pp 770–8

[87] Lawhern V J, Solon A J, Waytowich N R, Gordon S M, Hung C P and Lance B J 2018 EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces *J. Neural Eng.* **15** 056013

[88] Zheng W-L, Dong B-N and Lu B-L 2014 Multimodal emotion recognition using EEG and eye tracking data *2014 36th Annual Int. Conf. IEEE Engineering in Medicine and Biology Society* pp 5040–3

[89] Lan Z, Muller-Putz G R, Wang L, Liu Y, Sourina O and Scherer R 2016 Using support vector regression to estimate valence level from EEG *2016 IEEE Int. Conf. on Systems, Man and Cybernetics (SMC) (Budapest, Hungary)* (IEEE) pp 002558–63

[90] Lan Z, Sourina O, Wang L and Liu Y 2016 Real-time EEG-based emotion monitoring using stable features *Vis. Comput.* **32** 347–58

[91] Zheng W-L, Liu W, Lu Y, Lu B-L and Cichocki A 2019 Emotionmeter: a multimodal framework for recognizing human emotions *IEEE Trans. Cybern.* **49** 1110–22

[92] Liu W, Qiu J-L, Zheng W-L and Lu B-L 2022 Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition *IEEE Trans. Cogn. Dev. Syst.* **14** 715–29

[93] Hjorth B 1970 EEG analysis based on time domain properties *Electroencephalogr. Clin. Neurophysiol.* **29** 306–10

[94] Duan R-N, Zhu J-Y and Lu B-L 2013 Differential entropy feature for EEG-based emotion classification *2013 6th Int. IEEE/EMBS Conf. on Neural Engineering (NER)* pp 81–84

[95] Ng W B, Saidatul A, Chong Y F and Ibrahim Z 2019 PSD-based features extraction for EEG signal during typing task *IOP Conf. Ser.: Mater. Sci. Eng.* **557** 012032

[96] Stone N J 2000 Exploring the relationship between calibration and self-regulated learning *Educ. Psychol. Rev.* **12** 437–75

[97] Eva K W, Cunnington J P W, Reiter H I, Keane D R and Norman G R 2004 How can I know what I don't know? Poor self assessment in a well-defined domain *Adv. Health Sci. Educ.* **9** 211–24

[98] Karakaş S 2020 A review of theta oscillation and its functional correlates *Int. J. Psychophysiol.* **157** 82–99

[99] Fernández T, Harmony T, Rodríguez M, Bernal J, Silva J, Reyes A and Marosi E 1995 EEG activation patterns during the performance of tasks involving different components of mental calculation *Electroencephalogr. Clin. Neurophysiol.* **94** 175–82

[100] Picard R W, Papert S, Bender W, Blumberg B, Breazeal C, Cavallo D, Machover T, Resnick M, Roy D and Strohecker C 2004 Affective learning—a manifesto *BT Technol. J.* **22** 253–69

[101] Woolf B, Burleson W, Arroyo I, Dragon T, Cooper D and Picard R 2009 Affect-aware tutors: recognising and responding to student affect *Int. J. Learn. Technol.* **4** 129–64

[102] Rajendran R, Iyer S and Murthy S 2019 Personalized affective feedback to address students' frustration in ITS *IEEE Trans. Learn. Technol.* **12** 87–97

[103] Holmqvist K, Nyström M, Andersson R, Dewhurst R, Jarodzka H and van de Weijer J 2011 *Eye Tracking: A Comprehensive Guide to Methods and Measures* (Oxford: Oxford University Press)

[104] Kowler E 2011 Eye movements: the past 25 years *Vis. Res.* **51** 1457–83

[105] Lai M-L *et al* 2013 A review of using eye-tracking technology in exploring learning from 2000 to 2012 *Educ. Res. Rev.* **10** 90–115

[106] Alemdag E and Cagiltay K 2018 A systematic review of eye tracking research on multimedia learning *Comput. Educ.* **125** 413–28

[107] Mikhailenko M, Maksimenko N and Kurushkin M 2022 Eye-tracking in immersive virtual reality for education: a review of the current progress and applications *Front. Educ.* **7** 697032

[108] Stull A T, Fiorella L and Mayer R E 2018 An eye-tracking analysis of instructor presence in video lectures *Comput. Hum. Behav.* **88** 263–72