applied sciences

MDPI

Article

A Fusion Framework for Confusion Analysis in Learning Based on EEG Signals

Chenlong Zhang, Jian He, Yu Liang *, Zaitian Wang and Xiaoyang Xie

Faculty of Information, Beijing University of Technology, Beijing 100124, China; zhangchenlong@emails.bjut.edu.cn (C.Z.); jianhee@bjut.edu.cn (J.H.); wangzaitian@emails.bjut.edu.cn (Z.W.); xiexiaoyang@bjut.edu.cn (X.X.)

* Correspondence: yuliang@bjut.edu.cn

Abstract: Human–computer interaction (HCI) plays a significant role in modern education, and emotion recognition is essential in the field of HCI. The potential of emotion recognition in education remains to be explored. Confusion is the primary cognitive emotion during learning and significantly affects student engagement. Recent studies show that electroencephalogram (EEG) signals, obtained through electrodes placed on the scalp, are valuable for studying brain activity and identifying emotions. In this paper, we propose a fusion framework for confusion analysis in learning based on EEG signals, combining feature extraction and temporal self-attention. This framework capitalizes on the strengths of traditional feature extraction and deep-learning techniques, integrating local time-frequency features and global representation capabilities. We acquire localized time-frequency features by partitioning EEG samples into time slices and extracting Power Spectral Density (PSD) features. We introduce the Transformer architecture to capture the comprehensive EEG characteristics and utilize a multi-head self-attention mechanism to extract the global dependencies among the time slices. Subsequently, we employ a classification module based on a fully connected layer to classify confusion emotions accurately. To assess the effectiveness of our method in the educational cognitive domain, we conduct thorough experiments on a public dataset CAL, designed for confusion analysis during the learning process. In both subject-dependent and subject-independent experiments, our method attained an accuracy/F1 score of 90.94%/0.94 and 66.08%/0.65 for the binary classification task and an accuracy/F1 score of 87.59%/0.87 and 41.28%/0.41 for the four-class classification task. It demonstrated superior performance and stronger generalization capabilities than traditional machine learning classifiers and end-to-end methods. The evidence demonstrates that our proposed framework is effective and feasible in recognizing cognitive emotions.

Keywords: human–computer interaction; electroencephalographic; emotion recognition; confusion analysis; self-attention

Citation: Zhang, C.; He, J.; Liang, Y.; Wang, Z.; Xie, X. A Fusion Framework for Confusion Analysis in Learning Based on EEG Signals. *Appl. Sci.* **2023**, *13*, 12832. https://doi.org/10.3390/app132312832

Academic Editor: João M. F. Rodrigues

Received: 4 November 2023
Revised: 24 November 2023
Accepted: 28 November 2023
Published: 29 November 2023

1. Introduction

In modern education, human–computer interaction (HCI) plays a crucial role, with emotion recognition being particularly significant in the field of HCI. By accurately identifying and understanding students' emotional states, educational systems can better respond to their needs and provide personalized support. Emotion recognition technology can assist educators in determining whether students are experiencing confusion, frustration, or focus during the learning process, enabling timely adoption of appropriate teaching strategies and supportive measures [1–3]. Therefore, the importance of emotion recognition in HCI and education is self-evident. It optimizes the teaching process, enhances learning outcomes, and provides students with more personalized support and guidance. Confusion is more common than other emotions in the learning process [4–6]. Although confusion is an unpleasant emotion, addressing confusion during controllable periods has been shown to be beneficial for learning [7–9], as it promotes active student engagement in learning

activities. However, research on learning confusion is still in its early stages and requires further exploration.

Electroencephalography (EEG) is considered a physiological indicator of the aggregated electrical activity of neurons in the human brain's cortex. EEG is employed to record such activities and, compared to non-physiological indicators like facial expressions and gestures, offers a relatively objective assessment of emotions, making it a reliable tool for emotion recognition [10].

Traditionally, the classification of EEG signals relies on manual feature extractors and machine learning classifiers [11], such as Naive Bayes, SVM, and Random Forest. Although deep-learning architectures are a more recent introduction, they have consistently improved performance [12]. Convolutional Neural Networks (CNNs) and Long Short-Term Memory Networks (LSTMs) are the primary architectures employed [13]. However, employing CNNs for feature extraction primarily focuses on local aspects, hindering temporal information perception. Although LSTM-based approaches exhibit commendable performance, they also struggle with global temporal representation. Various attempts with end-to-end hybrid networks [14] have been made. However, these endeavors have resulted in models with excessively intricate architectures, leading to sluggish convergence rates or even failures to converge. Furthermore, end-to-end methodologies lack the advantages of conventional feature extraction methods in representing EEG signals. The Transformer [15] has showcased its formidable capabilities in natural language processing (NLP), owing to its significant advantage in comprehending global semantics. However, its application in EEG systems is still an area that requires further exploration.

In light of the limitations of end-to-end network models, as well as the disadvantages of CNN and LSTM, and considering the advantages of traditional feature extraction methods and the Transformer network structure, this paper proposes a fusion framework combining feature extraction and self-attention mechanism. Precisely, The EEG signals are first sliced, and frequency-domain features are extracted. These features are then tokenized into temporal tokens. Subsequently, the self-attention mechanism of the Transformer encoder layer is employed to capture temporal correlations. Finally, the extracted features are integrated using a fully connected layer to derive classification results, subsequently subjected to confusion analysis. We summarize the contributions of this paper as follows:

1. We present a fusion framework that integrates the strengths of traditional feature extraction and deep learning to analyze confusion during the learning process. This framework enables targeted guidance by assessing the cognitive level of students.
2. By harnessing the robust capabilities of the multi-head self-attention mechanism, we capture global contextual representations of long EEG segments, which proves beneficial for predicting confusion emotions.
3. In both subject-dependent and subject-independent experiments, we compare our framework with traditional machine learning classifiers and end-to-end methods. The experimental results demonstrate the superiority of our framework.

The rest of this paper is organized as follows: Section 2 presents the related work about the topic. Section 3 describes the methods. Section 4 presents the experiments and discusses the results. Section 5 describes the conclusion and future research directions.

2. Related Work

Confusion in learning refers to feeling perplexed or uncertain while absorbing knowledge or solving problems. Given its shared attributes with emotions, it is a nascent study area, primarily exploring confusion's classification as an emotion or affective state. Confusion is deemed a cognitive emotion, indicating a state of cognitive imbalance [9,16]. Individuals are encouraged to introspect and deliberate upon the material to redress this imbalance and facilitate progress, enabling a more profound comprehension. Consequently, when confused, individuals tend to activate profound cognitive processes to pursue enhanced learning outcomes. The investigation into confusion within the learning context remains in its preliminary stages.

Using EEG to recognize human emotions during various activities, including learning, is an area currently being explored. Recent research has focused on using electroencephalography to study cognitive states and emotions for educational purposes. These studies focus on attention or engagement [17,18], cognitive load, and some basic emotions such as happiness and fear. For example, researchers [19] used an EEG-based brain-computer interface (BCI) to record EEG in the FP1 region to track changes in attention. By utilizing visual and auditory cues, such as rhythmic hand raising, adaptive proxy robots can help students shift their attention when their attention falls below a preset threshold. The research results indicate that this BCI can improve learning performance.

Most traditional EEG-based classification methods rely on two steps: feature extraction and classification, and emotion classification is no exception. Many researchers have focused on exploring effective EEG features for classification, and the advancement of machine learning methods and technologies has significantly contributed to the development of these traditional methods. There have been attempts using the Common Spatial Pattern (CSP) algorithm [20], such as the FBCSP algorithm [21], which filters signals through filter banks, computes CSP energy features for each signal output through time filters, and then selects and classifies these features. Despite enhancements to the original CSP method, these techniques solely focus on analyzing the CSP energy dimension, disregarding the incorporation of temporal contextual information. Kaneshiro et al. [11] proposed Principal Component Analysis (PCA), extracting feature vectors of specific sizes from minimally preprocessed EEG signals, followed by training a classifier based on Linear Discriminant Analysis (LDA). Karimi-Rouzbahani et al. [22] explored the discriminative power of many statistical and mathematical features, and their experiments on three datasets showed that multi-valued features like wavelet coefficients and the theta frequency band performed better. Zheng et al. [23] investigated the pivotal frequency bands and channels of multichannel EEG data in emotion recognition. Jensen & Tesche [24] and Bashivan et al. [25] demonstrated through experiments that cortical oscillatory activity associated with memory operations primarily exists in the theta (4–7 Hz), alpha (8–13 Hz), and beta (13–30 Hz) frequency bands. The studies above utilize traditional machine learning classifiers to explore critical frequency bands and channels; nevertheless, traditional machine learning classifiers do not demonstrate any performance advantages. In addition, separately optimizing feature extraction and classifier could potentially result in suboptimal global optimization.

Compared to traditional methods, end-to-end deep networks eliminate the need for manual feature extraction. For most EEG applications, it has been observed that shallow models yield good results, while deep models might lead to performance degradation [12,13]. Especially for classification based on CNNs, despite the shallow architectures of CNNs with few parameters, they have been widely utilized: DeepConvNet [12], EEGNet [26], ResNet [27], and other variants [28]. However, due to the limitations imposed by kernel size, CNNs can learn features with local receptive fields. However, they cannot capture the crucial long-term dependencies for time series analysis. Furthermore, Recurrent neural networks(RNNs) and long short-term memory(LSTM) are introduced to capture the temporal features of EEG classification [29,30]. However, these models cannot be trained in parallel, and the dependencies calculated by hidden states quickly vanish after a few time steps, making it challenging to capture global temporal dependencies. Moreover, end-to-end methods insist on utilizing deep networks to learn from raw signals, often overlooking the advantages of manual feature extraction, and complex networks can lead to difficulties in model convergence.

3. Methods

Transformer [15] is an emerging neural network architecture that originated in machine translation tasks. In recent years, it has gained remarkable prominence in natural language processing. However, its application to emotion recognition based on EEG data remains an area requiring further research. In this paper, we combine feature extraction with Transformer for EEG classification. Drawing on the idea of Transformer, we first extract

local temporal and frequency features and then adopt the self-attention mechanism to capture global temporal features.

The overview of the proposed framework is depicted in Figure 1. It comprises three modules: preprocessing, multi-head self-attention, and a fully connected classifier. In the preprocessing module, the noise and artifacts of EEG signals are filtered out. Then, the temporal and frequency-domain information, encapsulating crucial local features, is extracted. Next, the multi-head self-attention module extracts long-term features by learning the global correlations between different temporal positions. Finally, utilizing the features extracted in the previous steps, which encapsulate spatial, frequency, and temporal information, a classifier composed of fully connected layers is adopted to output the classification results.

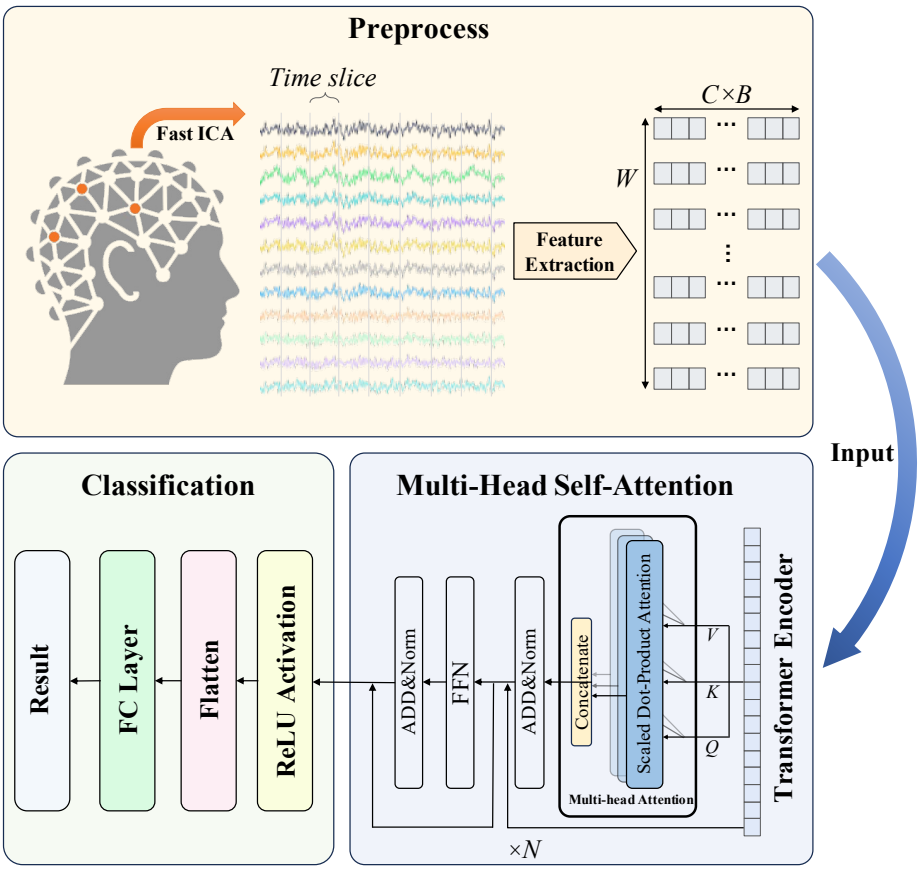


Figure 1. Overview of the proposed framework.

3.1. Preprocessing

Since the majority of EEG signals are concentrated within the range of 1 Hz to 50 Hz, a bandpass filter with a range of 1 Hz to 50 Hz was selected. This filtering procedure serves a dual purpose, eliminating low-frequency baseline drift, electrocardiographic (ECG) interference, and other high-frequency noise while effectively removing the most prominent power line interference (typically 50 Hz in China). Furthermore, EEG signals overlap with the electrooculogram (EOG) and electromyogram (EMG) signals in the frequency band. Therefore, relying solely on a single bandpass filter is insufficient to eliminate interference from EOG and EMG. This study adopted the fast, independent component analysis (fast ICA) to eliminate artifacts from EOG and EMG.

EEG comprises multiple time series corresponding to different spatial positions on the cerebral cortex where different electrodes are located on the collection device. Like audio signals, frequency-domain features are the most salient features. Thus, the spectrogram of the signals is typically employed for analysis. In frequency-domain analysis methods, Power Spectral Density (PSD) analysis is a typically adopted method, and most previous

studies have used this method to investigate epilepsy and hypnosis [31,32]. This method extracts frequency features that effectively detect cognitive and motor tasks [33]. Moreover, the PSD method consistently exhibits the highest robustness and effectiveness in extracting distinctive spectral patterns to differentiate motor imagery EEG signals accurately [34].

In this paper, Welch's method is used to extract the power spectrum features of EEG signals. The data sequence is applied to data windowing, producing modified periodograms [35]. For the EEG signal $x(n)$ of a certain channel, first divide it into L segments, with each segment having M sampling points, then the i-th small segment $x_i(n)$ can be denoted as:

$$x_i(n) = x(m + iM), 0 \leqslant m < M, 0 \leqslant i < L \tag{1}$$

take iD to be the point of start of the i th sequence. Then, L of length $2M$ represents data segments that are formed. The resulting output periodograms give:

$$\overset{\approx(i)}{P_{xx}}(f) = \frac{1}{MU}\left|\sum_{n=0}^{M-1} x_i(n)w(n)e^{-j2\pi fn}\right|^2 \tag{2}$$

Here, in the window function, U gives the normalization factor of the power and is chosen such that:

$$U = \frac{1}{M}\sum_{n=0}^{M-1} w^2(n) \tag{3}$$

where $w(n)$ is the window function. The average of these modified periodograms gives Welch's power spectrum as follows:

$$P_{xx}^{W} = \frac{1}{L}\sum_{i=0}^{L-1} \overset{\approx(i)}{P_{xx}}(f) \tag{4}$$

EEG signals predominantly reside within the 1 Hz to 50 Hz range, categorized into five frequency bands, depicted in Table 1. In this paper, the above method is applied to extract the PSD features of all channels and five frequency bands of EEG signals. Instead of manually selecting several of the five frequency bands for PSD feature extraction, the key features are extracted with the powerful ability of the transformer deep-learning network architecture. We divided each sample into 0.25 s slices. The sum of values within the five frequency bands is computed for each time slice and used as a separate measurement for each channel. This approach ensures that the extracted features encompass both frequency-domain and temporal information. PSD features for each time slice are extracted separately. This results in a sample with dimensions of $[W, B \times C]$, where W is the number of time slices, B is the number of frequency bands, and C is the number of channels. Figure 1 illustrates the above process.

3.2. Multi-Head Self-Attention

Due to the continuous nature of neural activity, the context-dependent representation between different time segments would contribute to EEG classification. This module uses self-attention to learn global temporal information of EEG features. The multi-head self-attention mechanism enables the model to attend to information from different representation subspaces from various channels. Each self-attention head, $h \in [1, 2, \ldots, H]$, with H being the total number of heads, relies on Q_c (queries), K_c (keys), V_c (values) vectors for token assessment. Within the features of each time segment, the frequency band features of different channels are concatenated sequentially. Given this, H is set to the number of channels, so The time slice representations are projected to latent representations of $Q_c, K_c, V_c \in \mathbb{R}^{1\times B}$, where $c \in [1, 2, \ldots, C]$. The "queries-keys" pair aims to map the key slices to the query slices based on their in-between representational similarity, calculated as their scaled dot-product, followed by SoftMax operation [15]. The resultant matrix is again multiplied with V_c to calculate the representational context as the aggregation of

the self-attentional interactions. For a set of C slices and a single self-attention head h, the representational context is calculated as

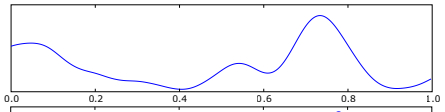
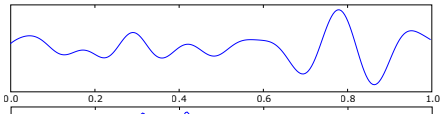
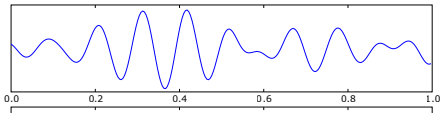
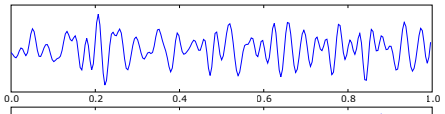
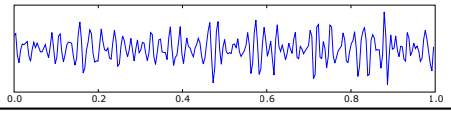
$$Head(Q_c, K_c, V_c)_h = Softmax(\frac{Q_c \cdot K_c^{Transpose}}{\sqrt{B}}) \cdot V_c \quad (5)$$

Multiple projections of Q_c, K_c, and V_c calculate the respective self-attention heads, and their outputs become concatenated to form the aggregate of multiple heads, as given by

$$\begin{aligned} &MultiHeadAttention(Q_c^{[1,2,\ldots,H]}, K_c^{[1,2,\ldots,H]}, V_c^{[1,2,\ldots,H]}) \\ &= Concat(Head_1, Head_2, \ldots, Head_H) \end{aligned} \quad (6)$$

In this module, N multi-head self-attention layers are employed. Finally, the output vector of this module is flattened, and a fully connected layer is utilized as the classifier to obtain EEG classification results.

Table 1. Five frequency bands of EEG signals.

Bands	Frequencies	States	Examples
delta	1–4 Hz	Sleep and dreaming	
theta	4–8 Hz	Deep relaxation or meditative states	
alpha	8–14 Hz	Resting or relaxed	
beta	14–31 Hz	Alert, active mind	
gamma	31–50 Hz	Intense focus, problem solving	

4. Experiments and Discussions

4.1. Dataset

We utilize a publicly available dataset called CAL, designed explicitly for confusion analysis in learning. The CAL dataset is first used in [1]. It focuses on cognitive emotions during the learning process, including four categories (confused, non-confused, guess, and think-right). Raven's Progressive Matrices [36] is employed as confusion stimuli to design the experiment. A total of 25 subjects participated in this experiment. Subjects watch ten scene pictures, each of which lasts 10 s. Next, they view and perform 48 tests, each lasting a maximum of 15 s. There are 23 subjects' data obtained because the unexpected equipment problem caused a failed collection for two persons. The participants evaluate their level of confusion for each test item at the end of the trials. OpenBCI is employed as the EEG collector, which has eight channels (Fp1, Fp2, C3, C4, T5, T6, O1, O2) and a 250 Hz sampling rate, depicted in Figure 2. Table 2 summarizes the relevant information of the CAL dataset. Each trial is segmented with a non-overlapped three-second time window. Each segment is regarded as one data sample during the model training.

Table 2. Summary of information on the CAL dataset.

Emotion Categories	Emotion Stimuli	#Subjects	#Channels	Sampling Rate
confused, non-confused, think-right, guess	tests	23 male/female: 12/11	8	250 Hz

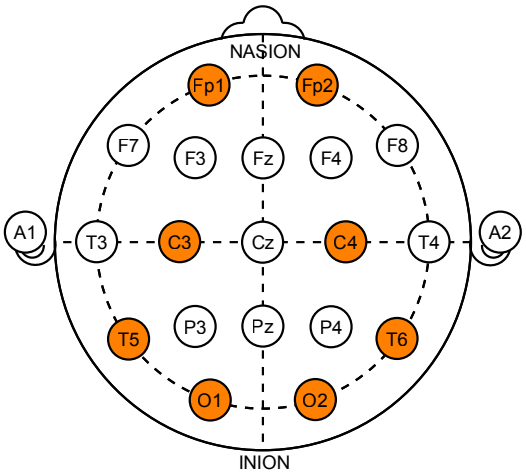


Figure 2. The eight channel locations (colored orange) of the CAL dataset, using the international 10–20 system.

4.2. Experiment Settings

We follow the settings outlined in [1] and conduct experiments involving binary classification (confused and non-confused) and four classification (confused, non-confused, guess, and think-right). In addition, we consider the following two scenarios to validate the proposed methods.

(1) subject-dependent: the data are trained across multiple subjects in the subject-dependent experiments. Specifically, 70% of the EEG data from all experiments for each participant are allocated as the training set, while the remaining 30% serve as the testing set.
(2) subject-independent: In the subject-independent model, the experiment emphasizes the differences between different subjects to test the method's generalization ability. Specifically, EEG data are divided into a cross-subject validation set with a split of 70%/30%, where the data from 16 subjects is used for training, and the data from the remaining seven subjects is used for testing.

To evaluate the classification performance of various methods, we consider the outcomes of conventional machine learning classifiers (Naive Bayes, SVM, and Random Forest) based on PSD features as presented in [1], as well as end-to-end methods (LSTM [30], ResNet [37], and EEGNet [26]). Furthermore, we conduct experiments using our approach without extracting PSD features to explore the benefits of feature extraction.

When extracting the samples of each category, we set a sliding window of 4 s to segment the data according to the setting in [1,38,39] to increase the sample size, i.e., the experimental samples of 4 s of EEG data. At the same time, to solve the data imbalance problem, we set overlapping parts of different lengths: 0.25 s for confused, 0.75 s for non-confused, 0.5 s for think-right, and 0.75 s for guess.

In this paper, the MNE [40] library in Python is adopted for data preprocessing operations. Our method is implemented with the PyTorch framework in Python 3.8 with an NVIDIA Geforce 3090 GPU. Using the same hyperparameter settings, we fix random seeds to repeat the experiment for different methods. We train the model using Adam optimizer with a learning rate of $1 \times e^{-4}$. The Adam optimizer combines the benefits of Momentum and RMSprop and adaptively adjusts the learning rate. $1 \times e^{-4}$ is a common starting learning rate. During training, batch size and dropout rates are set to 32 and 0.5.

The batch size is 32, which is a relatively small value to improve the model's generalization performance. A dropout rate of 0.5 is chosen, a common regularization technique that can reduce overfitting.

We set the number of multi-head self-attention layers N to 6, the number of heads H to the number of channels with a value of 14, and the dimension of the feed Forward layer to 2048. The number of parameters of the model in this configuration is 1M.

4.3. Analysis of Results

Tables 3 and 4 present the experimental results of different methods in subject-dependent and subject-independent experiments on the CAL dataset.

Table 3. Subject-dependent experiment of different methods on CAL dataset (Acc/F1 score). "w/" for "with" and "w/o" for "without". The best results are highlighted in **bold**.

Methods	Binary Classification	Four Classification
Naive Bayes w/ PSD	57.30/0.52	69.72/0.69
SVM w/ PSD	67.43/0.67	48.10/0.33
Random Forest w/ PSD	69.72/0.69	37.72/0.26
EEGNet	72.02/0.71	49.81/0.43
LSTM	73.45/0.73	53.29/0.49
ResNet	80.61/0.80	73.10/0.73
Our method w/o PSD	85.53/0.85	85.99/0.86
Our method	**90.49/0.90**	**87.59/0.87**

Table 4. Subject-independent experiment of different methods on CAL dataset (Acc/F1 score). "w/" for "with" and "w/o" for "without". The best results are highlighted in **bold**.

Methods	Binary Classification	Four Classification
Naive Bayes w/ PSD	60.59/0.55	40.83/0.27
SVM w/ PSD	55.71/0.53	38.05/0.31
Random Forest w/ PSD	55.98/0.52	35.94/0.25
EEGNet	64.46/0.61	39.14/0.20
LSTM	61.25/0.59	40.53/0.24
ResNet	57.95/0.55	40.05/0.23
Our method w/o PSD	65.85/0.63	40.88/0.39
Our method	**66.08/0.65**	**41.28/0.41**

The binary classification results of the subject-dependent experiments demonstrate that our approach significantly improves accuracy by 33.19%, 23.06%, and 20.77%, respectively, compared to conventional machine learning methods (Naive Bayes, SVM, and Random Forest). Additionally, we can observe that end-to-end deep-learning methods based on CNN, ResNet, and EEGNet perform well (with accuracies of 80.61% and 72.02%, respectively), indicating strong feature extraction capabilities of CNN-based methods. However, due to the limited receptive field of CNN, it struggles to capture global features. In contrast, the Transformer architecture based on the self-attention mechanism excels at capturing global information. Experimental results further confirm the advantages of the Transformer architecture: our method, based on Transformer, achieves an average accuracy increase of 18.47% and 9.88% compared to the CNN-based EEGNet and ResNet. Furthermore, we can observe that the end-to-end LSTM network performs impressively in the binary classification tasks of the subject-dependent experiments, achieving an accuracy of 73.45%. However, when faced with ultra-long time series data such as physiological signals, the end-to-end LSTM still loses information during training, leading to the inability to capture global features. In contrast, our Transformer-based approach can capture global

temporal context information, resulting in a 17.04% improvement in accuracy compared to the end-to-end LSTM method.

In the subject-dependent experiments for the four-classification task, our method also exhibits strong performance with an accuracy of 87.59%. Compared to the best-performing traditional machine learning method, Naive Bayes, our method achieves an improvement of 17.87% in accuracy. Additionally, compared to the top-performing end-to-end method, ResNet, our method shows an accuracy improvement of 14.49%.

Moreover, in subject-independent experiments, our approach achieves the highest accuracy of 68.08% and 41.74% in binary and four-class classifications, respectively. Compared to other methods, our approach achieves an accuracy improvement of 1.62% in binary classification and 0.45% in four-class classification.

The F1 score, which is based on precision and recall, is also an important evaluation metric [41]. We present the F1 score of the binary and four-class classification tasks in Figures 3 and 4. Our approach achieved F1 scores of 0.90, 0.87, 0.65, and 0.41 in the experiments. Compared to the best-performing methods, our approach demonstrates improvements of 0.10, 0.14, 0.04, and 0.10, respectively.

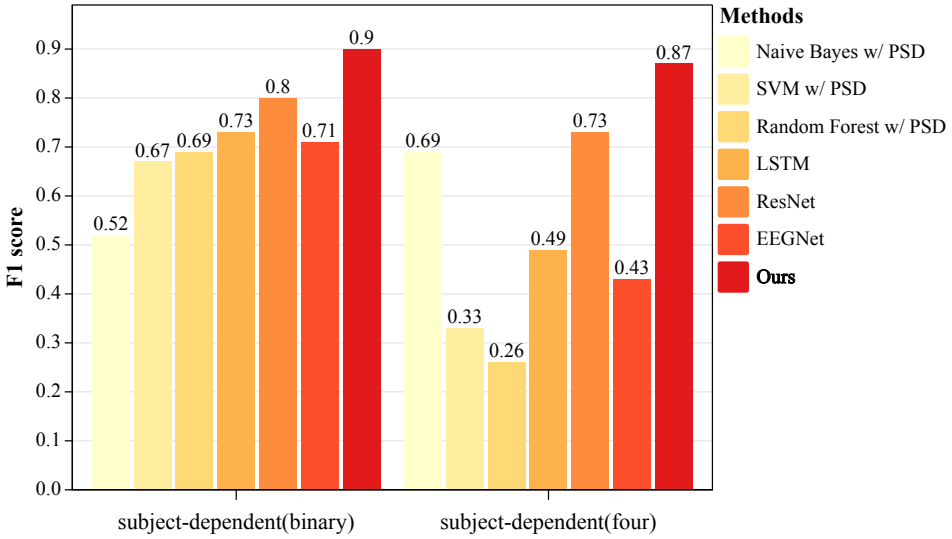


Figure 3. F1 score of different methods for the subject-dependent experiments.

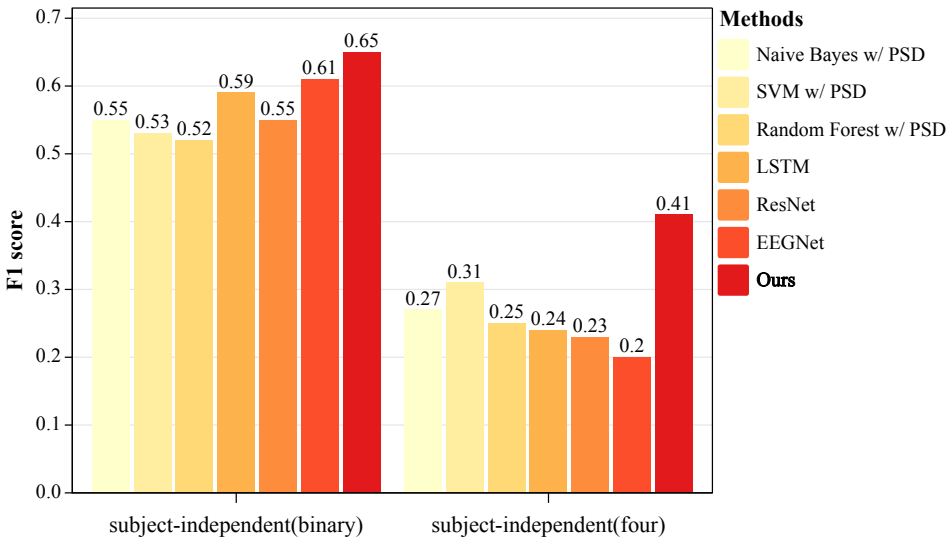


Figure 4. F1 score of different methods for the subject-independent experiments.

To evaluate the contribution of the feature extraction, we compare the performance of directly inputting the raw signal into the Transformer encoding layer without extracting PSD features. Comparing the last two rows of Tables 3 and 4, it is found that there is a decrease in performance. These ablation study results indicate PSD's effectiveness in representing EEG. The results also demonstrate the effectiveness of the feature extraction part in our proposed framework.

We also provide the accuracy and loss during the model training process. In Figure 5, the accuracy and loss curves are plotted to visualize the model's performance during training. The accuracy curve shows how well the model can correctly classify the data, while the loss curve indicates the error the model makes during training. We can see that both the accuracy and loss are improving as the number of epochs increases. It suggests that the model is learning and becoming more accurate over time. The convergence point, reached at around 2000 epochs, indicates that the model has achieved a stable performance and that further training may not lead to significant improvements. Overall, the provided accuracy and loss curves demonstrate that the model performs well and converges satisfactorily after approximately 2000 training epochs.

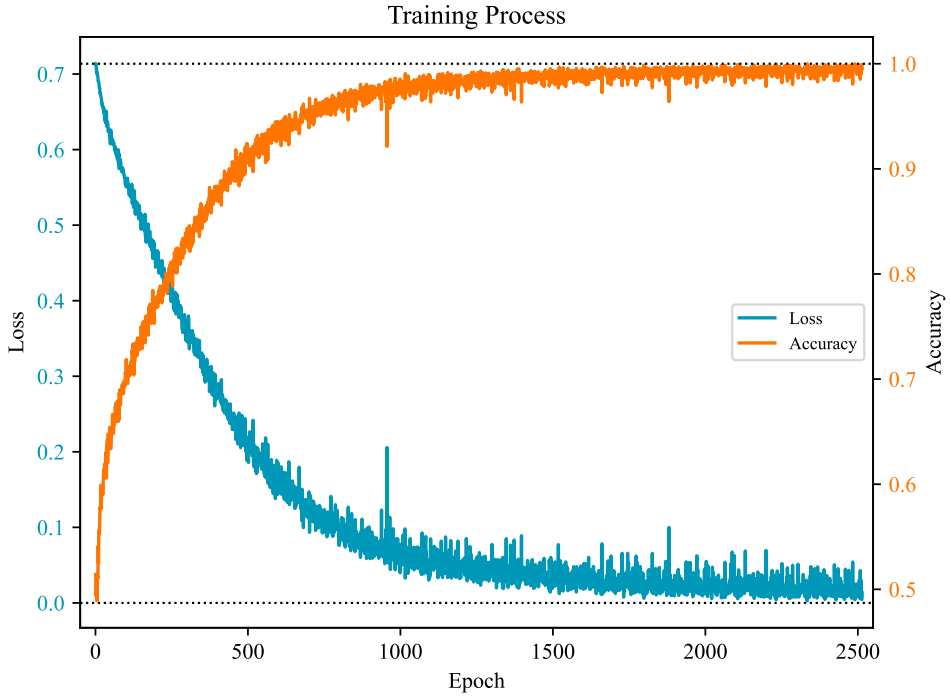


Figure 5. Accuracy and loss during the model training process.

The training time of the deep-learning model is also an important parameter. We compared the convergence time of all methods, as shown in Figure 6. Our study shows that our method has a significant advantage over the end-to-end deep-learning method in terms of training time. Our method offers a more efficient model that can save computing resources and time.

In addition to the faster training time, our method demonstrates better performance. The combination of faster training time and improved performance makes our method compelling for cognitive emotion identification. By reducing the computational burden and achieving better results, our method provides a practical and effective solution for this task. These advantages can have significant implications for real-world applications where efficiency and accuracy are crucial.

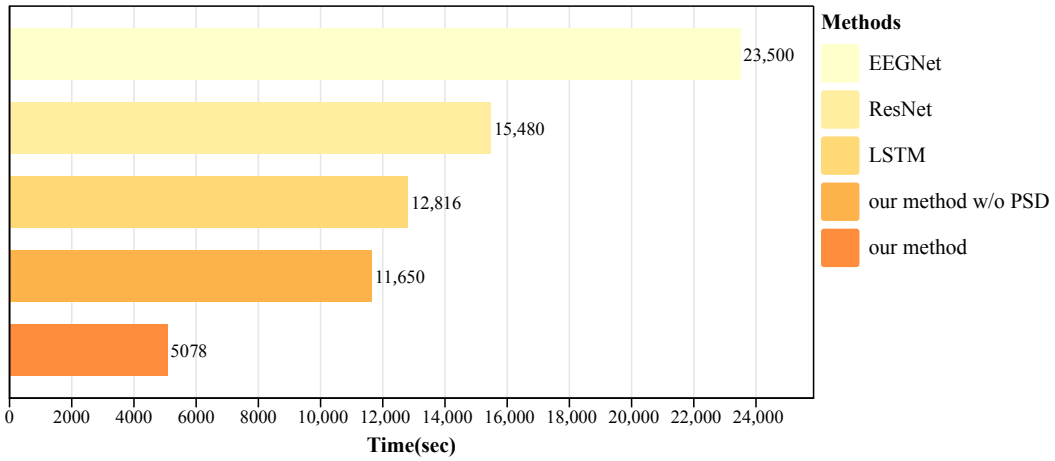


Figure 6. Training time for convergence of different methods.

4.4. Analysis of Confusion Matrix

We provide the confusion matrices of our approach for binary and four-class classification tasks for the subject-dependent and subject-independent experiments, as shown in Figures 7 and 8, respectively.

Our method performs well in identifying cognitive emotions during the subject-dependent experiment in binary and four-class setups. In the binary task, the model exhibits excellent discrimination between confused and non-confused emotions. In the four-class task, the recognition effect of confused emotions is slightly worse than that of the other three types of emotions. A total of 13% of the confused sample identified as think-right, suggesting that similar EEG patterns may be generated when subjects are confused or think-right. However, the identification performance is relatively poor in the subject-independent experiment due to the variations among different subjects. It is evident that in the case of binary classification, the model demonstrates good recognition of confused emotions among different subjects, but it struggles with identifying non-confused emotions. The same pattern can be observed in the four-class scenario, where the recognition performance is relatively better for confused emotions. This suggests that there is a certain degree of similarity in the EEG of different individuals when they are confused during learning, while there is a significant difference in the EEG of emotions other than confusion. These findings illustrate that addressing the differences among subjects poses a significant challenge.

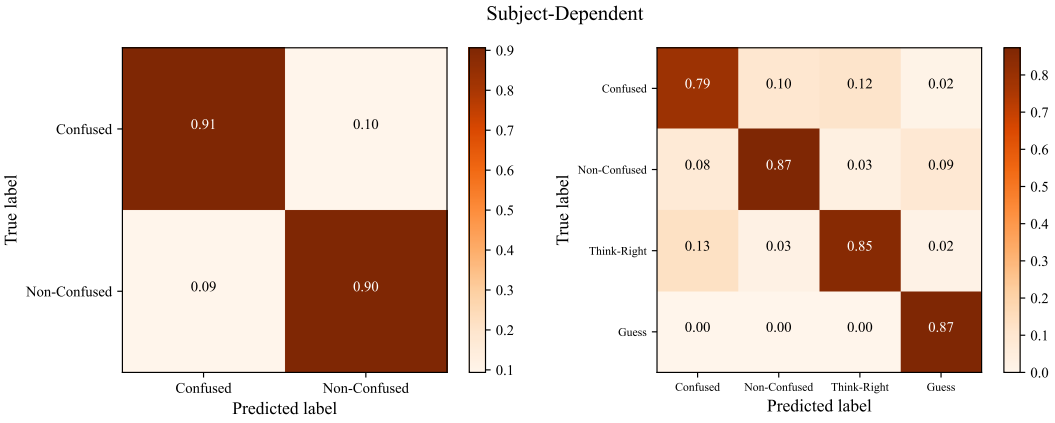


Figure 7. Confusion matrix of our method for the subject-dependent experiments.

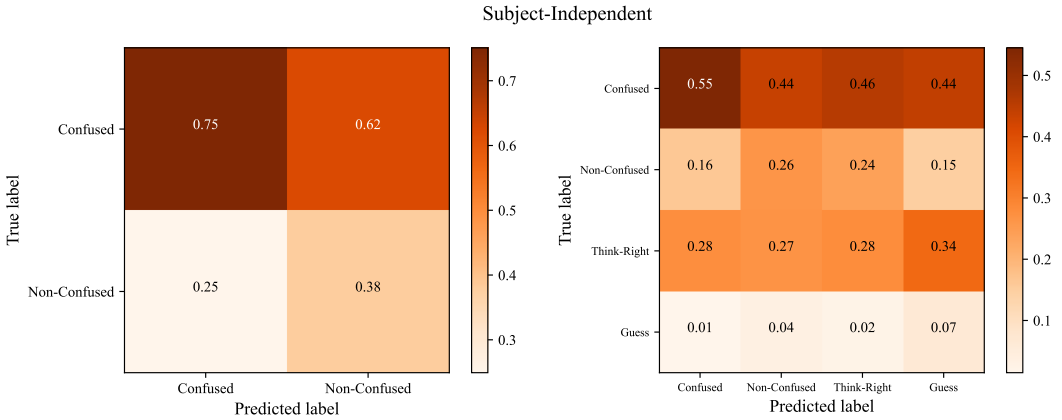


Figure 8. Confusion matrix of our method for the subject-independent experiments.

5. Conclusions and Future Work

In this work, we propose a fusion framework to analyze the presence of confusion in students during the learning process, aiming to expand the research on emotion recognition in education. This framework combines feature extraction and deep learning for analyzing confusion during learning. We first extract time-domain features with temporal information from EEG signals using a time-slicing approach and then utilize a multi-head self-attention mechanism to capture global-level temporal context representations. Extensive experiments conducted on the public dataset CAL demonstrate that our approach outperforms state-of-the-art methods in terms of performance and generalization ability. Furthermore, the effectiveness of our EEG representation approach can be extended to other physiological signal representations.

In future work, for the preprocessing module, we plan to extract multiple types of features to complement PSD features. For the subject-independent experiment, further research can focus on developing techniques that can effectively deal with the differences between different subjects to improve the universality and robustness of the model in identifying different individual cognitive emotions. It may involve collecting more diverse and representative data from more subjects. In addition, our work has certain limitations, including the small dataset size, the lack of additional physiological features, and any other potential constraints that may have influenced the findings. We plan to design experiments to collect multi-modal data (including EEG, ECG, facial expression, eye movement data, etc.) to build our dataset and explore emotion recognition in education by fusing multi-modal data.

Author Contributions: Conceptualization, C.Z. and Y.L.; methodology, C.Z.; visualization, Z.W.; validation, X.X.; formal analysis, X.X.; data curation, Z.W.; writing, C.Z. and Y.L.; supervision, J.H. and Y.L.; funding acquisition, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported in part by the Beijing Postdoctoral Research Foundation (No. 2023-22-97) and the State Key Laboratory of Software Development Environment (SKLSDE-2022KF-10).

Institutional Review Board Statement: The study was approved by the administrators of the public dataset used in the article.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The data presented in this study are openly available in CAL, MNE at https://iopscience.iop.org/article/10.1088/1741-2552/acbfe0 and https://www.frontiersin.org/articles/10.3389/fnins.2013.00267/full [1,40].

Acknowledgments: We thank the Beijing Engineering Research Center for IoT Software and Systems for providing the experimental environment and equipment.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Xu, T.; Wang, J.; Zhang, G.; Zhang, L.; Zhou, Y. Confused or not: Decoding brain activity and recognizing confusion in reasoning learning using EEG. *J. Neural Eng.* **2023**, *20*, 026018. [CrossRef]
2. Peng, T.; Liang, Y.; Wu, W.; Ren, J.; Pengrui, Z.; Pu, Y. CLGT: A graph transformer for student performance prediction in collaborative learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 15947–15954.
3. Liang, Y.; Peng, T.; Pu, Y.; Wu, W. HELP-DKT: An interpretable cognitive model of how students learn programming based on deep knowledge tracing. *Sci. Rep.* **2022**, *12*, 4012. [CrossRef]
4. Baker, R.S.; D'Mello, S.K.; Rodrigo, M.M.T.; Graesser, A.C. 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.* **2010**, *68*, 223–241. [CrossRef]
5. Han, Z.M.; Huang, C.Q.; Yu, J.H.; Tsai, C.C. Identifying patterns of epistemic emotions with respect to interactions in massive online open courses using deep learning and social network analysis. *Comput. Hum. Behav.* **2021**, *122*, 106843. [CrossRef]
6. Lehman, B.; Matthews, M.; D'Mello, S.; Person, N. What are you feeling? Investigating student affective states during expert human tutoring sessions. In Proceedings of the International Conference on Intelligent Tutoring Systems, Montreal, QC, Canada, 23–27 June 2008; Springer: Berlin/Heidelberg, Germany, 2008; pp. 50–59.
7. Lehman, B.; D'Mello, S.; Graesser, A. Confusion and complex learning during interactions with computer learning environments. *Internet High. Educ.* **2012**, *15*, 184–194. [CrossRef]
8. D'Mello, S.; Lehman, B.; Pekrun, R.; Graesser, A. Confusion can be beneficial for learning. *Learn. Instr.* **2014**, *29*, 153–170. [CrossRef]
9. Vogl, E.; Pekrun, R.; Murayama, K.; Loderer, K.; Schubert, S. Surprise, curiosity, and confusion promote knowledge exploration: Evidence for robust effects of epistemic emotions. *Front. Psychol.* **2019**, *10*, 2474. [CrossRef] [PubMed]
10. Gunes, H.; Piccardi, M. Bi-modal emotion recognition from expressive face and body gestures. *J. Netw. Comput. Appl.* **2007**, *30*, 1334–1345. [CrossRef]
11. Kaneshiro, B.; Perreau Guimaraes, M.; Kim, H.S.; Norcia, A.M.; Suppes, P. A representational similarity analysis of the dynamics of object processing using single-trial EEG classification. *PLoS ONE* **2015**, *10*, e0135697. [CrossRef]
12. Schirrmeister, R.T.; Springenberg, J.T.; Fiederer, L.D.J.; Glasstetter, M.; Eggensperger, K.; Tangermann, M.; Hutter, F.; Burgard, W.; Ball, T. Deep learning with convolutional neural networks for EEG decoding and visualization. *Hum. Brain Mapp.* **2017**, *38*, 5391–5420. [CrossRef] [PubMed]
13. Roy, Y.; Banville, H.; Albuquerque, I.; Gramfort, A.; Falk, T.H.; Faubert, J. Deep learning-based electroencephalography analysis: A systematic review. *J. Neural Eng.* **2019**, *16*, 051001. [CrossRef] [PubMed]
14. Yang, Y.; Wu, Q.; Qiu, M.; Wang, Y.; Chen, X. Emotion recognition from multi-channel EEG through parallel convolutional recurrent neural network. In Proceedings of the IEEE 2018 International Joint Conference on Neural Networks (IJCNN), Rio de Janeiro, Brazil, 8–13 July 2018; pp. 1–7.
15. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30* , 5998–6008.
16. Xu, T.; Zhou, Y.; Wang, Z.; Peng, Y. Learning emotions EEG-based recognition and brain activity: A survey study on BCI for intelligent tutoring system. *Procedia Comput. Sci.* **2018**, *130*, 376–382. [CrossRef]
17. Huang, J.; Yu, C.; Wang, Y.; Zhao, Y.; Liu, S.; Mo, C.; Liu, J.; Zhang, L.; Shi, Y. FOCUS: Enhancing children's engagement in reading by using contextual BCI training sessions. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Toronto, ON, Canada, 26 April–1 May 2014; pp. 1905–1908.
18. Xu, T.; Wang, X.; Wang, J.; Zhou, Y. From textbook to teacher: An adaptive intelligent tutoring system based on BCI. In Proceedings of the 2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), Guadalajara, Mexico , 1–5 November 2021; pp. 7621–7624.
19. Xu, J.; Zhong, B. Review on portable EEG technology in educational research. *Comput. Hum. Behav.* **2018**, *81*, 340–349. [CrossRef]
20. Ramoser, H.; Muller-Gerking, J.; Pfurtscheller, G. Optimal spatial filtering of single trial EEG during imagined hand movement. *IEEE Trans. Rehabil. Eng.* **2000**, *8*, 441–446. [CrossRef] [PubMed]
21. Ang, K.K.; Chin, Z.Y.; Wang, C.; Guan, C.; Zhang, H. Filter bank common spatial pattern algorithm on BCI competition IV datasets 2a and 2b. *Front. Neurosci.* **2012**, *6*, 39. [CrossRef] [PubMed]
22. Karimi-Rouzbahani, H.; Shahmohammadi, M.; Vahab, E.; Setayeshi, S.; Carlson, T. Temporal codes provide additional category-related information in object category decoding: A systematic comparison of informative EEG features. *bioRxiv* **2020**. . [CrossRef]
23. Zheng, W.L.; Lu, B.L. Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Trans. Auton. Ment. Dev.* **2015**, *7*, 162–175. [CrossRef]
24. Jensen, O.; Tesche, C.D. Frontal theta activity in humans increases with memory load in a working memory task. *Eur. J. Neurosci.* **2002**, *15*, 1395–1399. [CrossRef]
25. Bashivan, P.; Bidelman, G.M.; Yeasin, M. Spectrotemporal dynamics of the EEG during working memory encoding and maintenance predicts individual behavioral capacity. *Eur. J. Neurosci.* **2014**, *40*, 3774–3784. [CrossRef]
26. Lawhern, V.J.; Solon, A.J.; Waytowich, N.R.; Gordon, S.M.; Hung, C.P.; Lance, B.J. EEGNet: A compact convolutional neural network for EEG-based brain–computer interfaces. *J. Neural Eng.* **2018**, *15*, 056013. [CrossRef]

27. Tian, T.; Wang, L.; Luo, M.; Sun, Y.; Liu, X. ResNet-50 based technique for EEG image characterization due to varying environmental stimuli. *Comput. Methods Programs Biomed.* **2022**, *225*, 107092. [CrossRef]
28. Kalafatovich, J.; Lee, M.; Lee, S.W. Decoding visual recognition of objects from eeg signals based on attention-driven convolutional neural network. In Proceedings of the 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Toronto, ON, Canada, 11–14 October 2020; pp. 2985–2990.
29. Chowdary, M.K.; Anitha, J.; Hemanth, D.J. Emotion recognition from EEG signals using recurrent neural networks. *Electronics* **2022**, *11*, 2387. [CrossRef]
30. Lu, P. Human emotion recognition based on multi-channel EEG signals using LSTM neural network. In Proceedings of the IEEE 2022 Prognostics and Health Management Conference (PHM-2022 London), London, UK, 27–29 May 2022; pp. 303–308.
31. Fraiwan, L.; Lweesy, K.; Khasawneh, N.; Wenz, H.; Dickhaus, H. Automated sleep stage identification system based on time–frequency analysis of a single EEG channel and random forest classifier. *Comput. Methods Programs Biomed.* **2012**, *108*, 10–19. [CrossRef] [PubMed]
32. Deivanayagi, S.; Manivannan, M.; Fernandez, P. Spectral analysis of EEG signals during hypnosis. *Int. J. Syst. Cybern. Informatics* **2007**, *4*, 75–80.
33. Brodu, N.; Lotte, F.; Lécuyer, A. Exploring two novel features for EEG-based brain–computer interfaces: Multifractal cumulants and predictive complexity. *Neurocomputing* **2012**, *79*, 87–94. [CrossRef]
34. Duan, L.; Zhong, H.; Miao, J.; Yang, Z.; Ma, W.; Zhang, X. A voting optimized strategy based on ELM for improving classification of motor imagery BCI data. *Cogn. Comput.* **2014**, *6*, 477–483. [CrossRef]
35. Faust, O.; Acharya, R.; Allen, A.R.; Lin, C. Analysis of EEG signals during epileptic and alcoholic states using AR modeling techniques. *Irbm* **2008**, *29*, 44–52. [CrossRef]
36. Raven, J. The Raven's progressive matrices: Change and stability over culture and time. *Cogn. Psychol.* **2000**, *41*, 1–48. [CrossRef]
37. Wang, P.; Guo, C.; Xie, S.; Qiao, X.; Mao, L.; Fu, X. EEG emotion recognition based on knowledge distillation optimized residual networks. In Proceedings of the 2022 IEEE 6th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), Beijing, China, 3–5 October 2022; pp. 574–581.
38. Zheng, W.L.; Liu, W.; Lu, Y.; Lu, B.L.; Cichocki, A. Emotionmeter: A multimodal framework for recognizing human emotions. *IEEE Trans. Cybern.* **2018**, *49*, 1110–1122. [CrossRef]
39. Liu, W.; Qiu, J.L.; Zheng, W.L.; Lu, B.L. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Trans. Cogn. Dev. Syst.* **2021**, *14*, 715–729. [CrossRef]
40. Gramfort, A.; Luessi, M.; Larson, E.; Engemann, D.A.; Strohmeier, D.; Brodbeck, C.; Goj, R.; Jas, M.; Brooks, T.; Parkkonen, L.; et al. MEG and EEG data analysis with MNE-Python. *Front. Neurosci.* **2013**, *7*, 267. [CrossRef] [PubMed]
41. Powers, D. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. *J. Mach. Learn. Technol.* **2011**, *2*, 37–63.